BRUMFIELD, TERESA E., Ph.D. Fidelity of Test Development Process within a National Science Grant. (2007)
Directed by Dr. Terry A. Ackerman. 369 pp.

In 2002, a math-science partnership (MSP) program was initiated by a national science grant. The purpose of the MSP program was to promote the development, implementation, and sustainability of promising partnerships among institutions of higher education, K-12 schools and school systems, as well as other important stakeholders. One of the funded projects included a teacher-scientist collaborative that instituted a professional development system to prepare teachers to use inquiry-based instructional modules.

The MSP program mandated evaluations of its funded projects. One of the teacher-scientist collaborative project's outcomes specifically focused on teacher and student science content and process skills. In order to provide annual evidence of progress and to measure the impact of the project's efforts, and because no appropriate science tests were available to measure improvements in content knowledge of participating teachers and their students, the project contracted for the development of science tests.

This dissertation focused on the *process* of test development within an evaluation and examined planned (i.e., expected) and actual (i.e., observed) test development, specifically concentrating on the factors that affected the actual test development process. Planned test development was defined as the process of creating tests according to the well-established test development procedures recommended by the AERA/APA/NCME

1999 *Standards for Educational and Psychological Testing.* Actual test development was defined as the process of creating tests as it actually took place.

Because case study provides an in-depth, longitudinal examination of an event (i.e., case) in a naturalistic setting, it was selected as the appropriate methodology to examine the difference between planned and actual test development. The case (or unit of analysis) was the test development task, a task that was bounded by the context in which it occurred—and over which this researcher had no control—and by time. The purpose for studying the case was to gain a more in-depth, holistic understanding of the real-life test development task that took place within a project evaluation context. In particular, this case study investigated how the actual test development process was affected by:

1. the national and state (i.e., NC) science standards,
2. the NSF's definition of "evidence" in a project evaluation,
3. the MSP project's understanding of the role of the to-be-developed tests in their project evaluation,
4. the MSP project's understanding of the test development process, and
5. the MSP project's participants (e.g., teacher item-writers and scientists).

From an investigation of this case, it was concluded that:

- constructing psychometrically sound tests within an evaluation is not easy,
- sufficient time and resources to construct such measures properly are seldom provided, and
- test construction—at least within an evaluation—is not routine and unproblematic.

Based upon the results from this case study, it was recommended that stakeholders (i.e., program managers, project directors, and evaluators) be familiar with the steps and standards used to develop psychometrically sound tests. Additionally, it was recommended that, for future research, a meta-analysis that examines *only* the test development process be conducted of all other MSP projects.

A second suggested future research area was to establish a protocol that provides a systematic means by which to examine an existing or proposed MSP project for alignment with state science standards. Such a protocol would be cost-effective in that demonstrated alignment with state science standards would enable projects to use existing state science assessments, which must be in place, according to *NCLB*, by the 2007-2008 school year, to demonstrate student achievement. In this way, project directors and evaluators, typically with limited familiarity with the steps and standards by which psychometrically sound assessments are created, would not be placed in the role of test developer.

FIDELITY OF TEST DEVELOPMENT PROCESS

WITHIN A NATIONAL SCIENCE GRANT

by

Teresa E. Brumfield

A Dissertation Submitted to
the Faculty of The Graduate School at
The University of North Carolina at Greensboro
in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Greensboro
2007

Approved by

_____
Committee Chair

To Adam

APPROVAL PAGE

This dissertation has been approved by the following committee of the Faculty of The Graduate Schools at The University of North Carolina at Greensboro.

Committee Chair _____

Committee Members _____

_____

_____

_____

_____
Date of Acceptance by Committee

_____
Date of Final Oral Examination

ACKNOWLEDGEMENTS

I want to acknowledge and thank my committee chair, Dr. Terry Ackerman, along with the other members of my committee, Deb Bartz, Drs. Kristin Bennett, Betty Epanchin, and John Willse, for their patient assistance and guidance. I also want to thank the team of scientists at Duke University with whom I had the privilege of working on this project.

I also want to acknowledge and thank God for His gracious enablement and encouragement through my husband (especially!) and through my family and friends to persevere to the end and to not give up.

TABLE OF CONTENTS

Page

Page

Page

LIST OF TABLES

Page

Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Due in part to the No Child Left Behind Act of 2001, accountability, alignment, and assessment are currently topics of interest in the field of education. This legislation, which aims to raise achievement for every child, mandates academic standards and assessments in reading/language arts and math for each of grades 3 through 8 and high school as well as academic standards and assessments in science for elementary, middle, and high schools. (20 U.S.C. §6301 et. seq.)

Essentially, the NCLB legislation holds state educational systems accountable to its stakeholders (e.g., parents, students, taxpayers) for one of the outcomes of this system—student academic achievement. In a similar fashion, funded programs (e.g., North Carolina Partnership in Science and Mathematics) are held accountable to their stakeholders, especially their funding agencies (e.g., National Science Foundation), to provide evidence that the program is doing what it set out to do. Assessments (i.e., tests) are one of the ways programs provide evidence that the program is meeting its proposed (and funded) outcomes (e.g., math achievement).

Educational decision-making is at the center of education testing. For instance, when an instructor assesses students' strengths and weaknesses, the instructor uses test results to *decide* what instructional objectives to pursue. When an instructor assesses

students' progress, the instructor uses test results to *decide* whether certain parts of the instructional program need to be altered. When an instructor assesses students to assign grades, the instructor uses students' performances to *decide* which students get which grades. Lastly, when an instructor uses pretest-to-posttest assessment results to indicate how effective an instructional sequence has been, the instructor is attempting to *decide* whether the instructional sequence needs to be revised. In fact, some believe instructors should never assess students without a clear understanding of what the decision is that will be informed by results from the assessment. (Popham, 1999).

Because educational assessment is used for educational decision-making, the more psychometrically sound an assessment, the more confident one can be in the decisions based on that assessment's results. In fact, it is the use of tests that is the primary focus of the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing.* These *Standards*, which promote the sound and ethical use of tests, provide criteria for the evaluation of tests, testing practices, and the effects of test use.

In early 2005, the acting director of the Center for Educational Research and Evaluation (CERE) at the University of North Carolina-Greensboro (UNCG) was contacted by the project director of a funded national science grant. The project director knew of CERE's expertise in psychometrics and educational testing as well as in program evaluation. He proposed to enter into a subcontract with CERE for the development of a series of elementary/middle school science tests that he needed in his project's evaluation.

The CERE director, in turn, approached this researcher, an employee of CERE, to head up this science test development project. This researcher was particularly suited for work on the project because, as a doctoral candidate, she had been trained in UNCG's Department of Educational Research Methodology in the field of psychometrics and evaluation; and she had worked on previous test development projects through the Center. Thus as test developer, this researcher was an active participant, rather than a passive observer, in the development of these science tests. (Stake, 1995).

With the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing* to guide a test developer, one would expect the process of developing assessments for a funded program to use to provide evidence to its funding agency that the program is meeting its targeted outcomes (e.g., increased science content knowledge) to be systematic, mostly uncomplicated, and problem-free. However, this researcher/test developer learned that what the process actually looked like in practice was quite different. Thus this dissertation focuses on the gap between theory (or what one would have expected to happen) and practice (or what actually happened) that arose in the process of this researcher developing a series of elementary grade science assessments for a funded National Science Foundation (NSF) project.

To provide a general overview of the context in which this dissertation takes place, this chapter begins with a short discussion of evaluation and the historical background of the NSF project for which the science assessments were developed. The chapter then presents the statement of the problem, the research objectives of this study, and its professional significance.

In its broadest sense, to evaluate means to (1) determine or set the value of or amount of; appraise: to evaluate property; (2) judge or determine the significance, worth, or quality ; assess: to evaluate the results of an experiment. (Retrieved 10/2/06 from Dictionary.com website: http://dictionary.reference.com/browse/evaluate.)

The Joint Committee on Standards for Education Evaluation, created in 1975, is a coalition of major professional associations concerned with the quality of evaluation. Widely recognized evaluation standards emanating from the Joint Committee include: *The Personnel Evaluation Standards*, originally published in 1988 with a draft second edition published in August 2006; *The Program Evaluation Standards, Second Edition*, published in 1994 by Sage Publications; and *The Student Evaluations Standards*, published in 2003. (Retrieved 9/30/06 from http://www.wmich.edu/evalctr/jc/ .)

The Joint Committee on Standards for Education Evaluation (1994) defines the terms *evaluation* as a " systematic investigation of the worth or merit of an object; e.g., a program, project, or instructional material" and *program evaluation* as an evaluation that assesses "activities that are funded for a defined period of time to perform a specified task [e.g.,] a three-day workshop on behavioral objectives, a two-year development effort, …." (p.208)

Rossi, Lipsey, and Freeman (2004) used evaluation in a more restricted sense, as program evaluation or interchangeably as evaluation research, which they define as "a social science activity directed at collecting, analyzing, interpreting, and communication information about the workings and effectiveness of social programs." (p. 2) The authors indicated some of the practical reasons for conducting evaluations: to aid in

decisions concerning whether programs should be continued, improved, expanded, or curtailed; to assess the utility of new programs and initiatives; to increase the effectiveness of program management and administration; and to satisfy the accountability requirements of program sponsors. In addition, evaluations may contribute to substantive and methodological social science knowledge.

Rossi, et al. (2004) stated that, at various times, policymakers, funding organizations, planners, program managers, taxpayers, or program clientele need to distinguish worthwhile social programs from ineffective ones and launch new programs or revise existing ones so as to achieve certain desirable results. To do so, they must seek answers to questions such as:

- What are the appropriate target populations for intervention?
- Is a particular intervention reaching its target population?
- Is the intervention being implemented well? Are the intended services being provided?
- Is the intervention effective in attaining the desired goals or benefits?
- Is the program cost reasonable in relation to its effectiveness and benefits?

The authors pointed out that answers to such questions are necessary not only for local or specialized programs, such as job training in a small town or a new mathematics curriculum for elementary schools, but also for broad national or state programs in such areas as health care, welfare, and educational reform.

For the purposes of this dissertation, evaluation is used as Rossi, et al. (2004), defined it--"a social science activity directed at collecting, analyzing, interpreting, and

communication information about the workings and effectiveness of social programs." (p. 2) The program of interest is the Math-Science Partnership (MSP) Program, sponsored by the National Science Foundation, that took place within the larger context of science education reform. Within the MSP Program are numerous projects; the project of interest in the context of this study is the Teachers and Scientists Collaborating (TASC) project at Duke University, and the evaluation question of interest is whether or not TASC had been effective in attaining its stated goals and objectives. The next section provides background information on science education reform, the Math-Science Partnership Program, and the particular MSP project, TASC.

Background

In January 2002, the President signed into effect Public Law 107-110—i.e., the No Child Left Behind Act of 2001, legislation designed "to close the achievement gap with accountability, flexibility, and choice, so that no child is left behind." (20 U.S.C. 6301 et seq.). Title II of Public Law 107-110, which addressed the preparation, training, and recruitment of high quality teachers and principals, specifically purposed in Part B—Mathematics and Science Partnerships—to improve students' academic achievement in the areas of mathematics and science. State educational agencies, institutions of higher education, local educational agencies, elementary schools, and secondary schools were encouraged to participate in programs that, among other things, would increase the subject matter knowledge and teaching skills of K-12 mathematics and science teachers through collaborations with scientists, mathematicians, and engineers.

In 2002, the National Science Foundation (NSF), in support of improving student outcomes in mathematics and science for all K-12 students, launched its Math and Science Partnership (MSP) program that promotes the development, implementation, and sustainability of promising partnerships among institutions of higher education, K-12 schools and school systems, as well as other important stakeholders. (Retrieved 12/14/05 from http://www.nsf.gov.) Under the MSP program, NSF awarded, in October 2002, a $5.3 million, five-year contract to Duke University. This MSP project—Teachers and Scientists Collaborating, or TASC—included Duke University Pratt School of Engineering, four (at the time of award) North Carolina school districts, the North Carolina Department of Public Instruction, and the North Carolina Science, Mathematics, and Technology Education Center. (NSF Award Abstract - #0227035).

Objectives of the TASC project included establishing a group of scientists who would provide ongoing teacher assistance in science content that was aligned with state/national standards, instituting a professional development system to prepare teachers to use inquiry-based instructional modules and to benefit from scientist resources, creating a fee-based lending library of inquiry-based modules available to teachers, and institutionalizing science education support. Through this project, which was expected to serve approximately 7,500 teachers and 353,000 students, TASC sought to: narrow achievement gaps for at-risk (i.e., below poverty level) and minority students; improve science and mathematics scores on state-mandated end-of-grade tests; and improve the quality of science teaching in participating school systems by increasing teachers' content knowledge, by teachers' use of inquiry-based teaching techniques, and by engaging

scientists to assist teachers in implementing science standards. (NSF Award Abstract - #0227035).

One of Duke University's responsibilities to NSF under this contract was to provide evidence of the strengths and weaknesses of the TASC project, thereby facilitating the MSP's understanding of what works. That is, an evaluation of the TASC project was required to guide the annual assessment of progress and to measure the impact of the project's efforts. (Program Solicitation NSF-02-061, 2002). One of the TASC project's evaluation activities specifically focused on teacher and student science content and process skills and called for the development of science tests to measure improvements in content knowledge of participating teachers and their students.

Statement of the Problem

Worthen, Sanders, and Fitzpatrick (1997) pointed out that tests are one method for collecting evaluative information and that for educational evaluators in particular, tests constitute a major source of information. The authors indicated that knowledge acquisition is frequently the primary objective of educational programs, and the acquisition of knowledge is generally measured by tests.

Whereas Worthen, et al. (1997) recognized the *use* of tests by educational evaluators, Rossi, Lipsey, and Freeman (2004) pointed out that, when measures must be developed to appraise a project's outcomes of interest, frequently, there is rarely sufficient time and resources to do this properly within the evaluation. These authors acknowledged that constructing such measures as questionnaires, attitude scales, and knowledge tests so that they measure what they are supposed to measure in a consistent

fashion is often not easy and that because of this, there are well-established measurement procedures to be followed. These procedures involve a number of technical considerations and generally require pilot testing, analysis, revision, and validation before a newly developed measure can be used with confidence.

Wolf and Cumming (2000) acknowledged that assessment and measurement consume large amounts of time in education and that formal assessment procedures are increasingly important in program evaluation and in public monitoring of education systems. However, they stated that the research literature in both psychometrics and performance assessment (i.e., assessment designed to measure demonstrated achievement rather than underlying traits) tends to treat actual test construction as unproblematic. In addition, they indicated that there is "remarkably little discussion in the academic literature" as to how an instrument actually gets developed.

As test developer for the MSP project evaluation, the goal was to create nine science tests—each test to be aligned with one of the science competency goals of the 2004 North Carolina Standard Course of Study for third, fifth, and eighth grade—according to the established measurement procedures of the field of psychometrics. As pointed out by Wolf and Cumming (2000), the tendency of psychometrics is to treat actual test construction a routine activity. These researchers, however, in their creation of a special assessment instrument learned otherwise, as did this researcher in the creation of these science tests.

This dissertation examines the incongruence between the planned (i.e., expected) test development process and the actual (i.e., observed) test development process and

suggests ways to reduce such inconsistencies for future research efforts. For the purpose of this study, planned test development is defined as the process of creating tests according to the well-established test development procedures recommended by the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing.* Actual test development is defined as the process of creating tests as it actually took place or, as phrased by Wolf & Cumming (2000) "the very messy grass roots of test development" (p. 211).

Objectives

The objectives of this dissertation are to (1) present the well-established test development procedures recommended by the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing,* including the analytical techniques used to develop these science tests; (2) address the difficulties that were encountered throughout the actual development of these tests; and (3) suggest ways to make planned test development and actual test development more congruent. The overall research question addressed is: How did the actual (i.e., observed) test development process differ from the planned (i.e., expected) test development process? More specifically, the factors that affected the development of these tests and how they affected the development of these tests were investigated.

The second chapter of the current research project, graphically depicted as four embedded rings (Figure 2 from Chapter Two), presents the situational context in which this project takes place. The discussions begins with current science education reform—the "big picture" context and the most general influence on the development of the

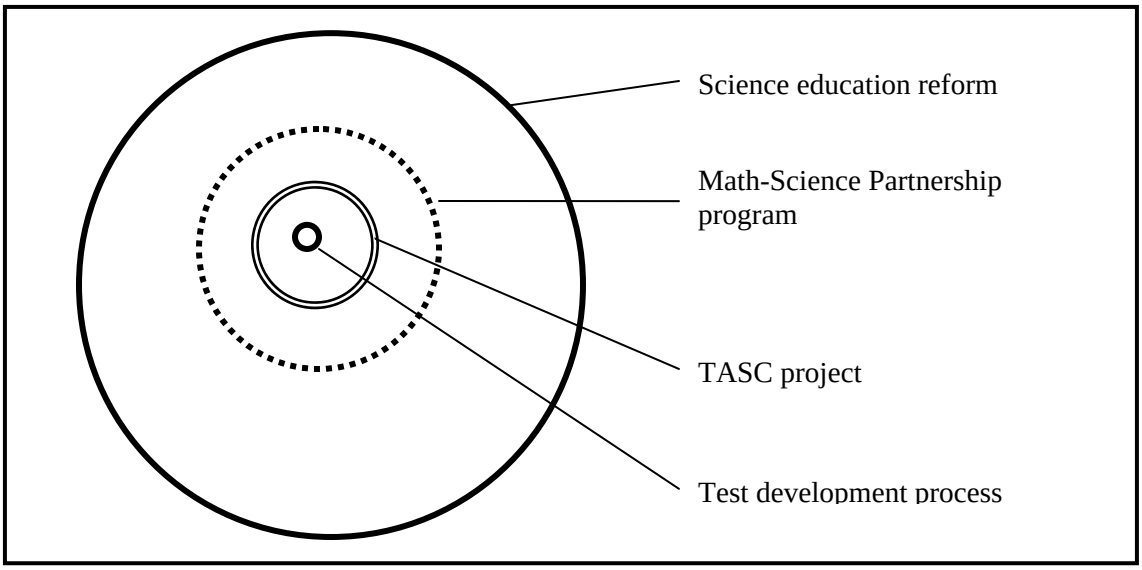


Figure 2 (from Chapter Two). Context of test development process

science tests. Drilling down to the next level of influence, the discussion moves to the National Science Foundation's Math-Science Partnership program, one of many programs within science education reform. Again drilling down, the discussion moves to the specific MSP project (again, one of many projects within the MSP program) of relevance to the current research—Teachers and Scientists Collaborating (TASC) project. Lastly, although there are many tasks that take place within a project's evaluation, the task of relevance to the present study is assessment—specifically, the creation of science tests to be used by project personnel to collect evidence concerning the effectiveness of the project. Thus, even though this study occurred within a project evaluation, it is the process of test development that is the focus of this dissertation.

Chapters Three, Four, and Five present the methodology (i.e., case study) used for this dissertation, the results, and conclusions and recommendations, respectively.

Professional Significance

The research literature does not lack discussion regarding planned test construction, i.e, *expected* test construction. For instance, Crocker and Algina (1986) and Allen and Yen (1979) presented a sequence of steps in the systematic approach to test construction. However, as stated previously by Wolf and Cumming (2000), the research literature lacks discussion as to how a test is actually constructed—that is, *observed* test construction. Thus one contribution of the current research project is to supplement existing research with a discussion of how a test actually gets developed, particularly within an evaluation context. A second contribution of the current research project is to make researchers aware of the potential pitfalls of actual test development in order that they may be proactive in creating tests as part of a program evaluation.

Lastly, by focusing on the test development task within a MSP project evaluation, this study informs not only TASC, but also the Math-Science Partnership Program, NSF, and ultimately, science education reform. Providing teachers with the *NSES*-based content and pedagogical knowledge is the goal of many science education professional development programs, including TASC. Evaluating these programs requires assessment instruments. However, as pointed out by Assessing Teacher Learning About Science Teaching—a MSP Research, Evaluation, and Technology Assistance project funded by NSF—a coherent set of tools, which currently does not exist, are needed by professional development providers to inform revisions to their program designs and implementations. While ATLAST is a MSP RETA *project* funded by NSF for the purpose of developing instruments in specific science content areas, this dissertation documented how various

"rings of influence" affected the test development *task* that took place within a MSP project evaluation, and how this influence on the test development task in turn affected the data collection process required by professional development providers not only to inform revisions to their programs but also to provide evidence to their funders of their program's effectiveness. While TASC is the "center of attention" for this dissertation, it is merely an example of what is, most likely, common practice in the evaluation of science education programs.

CHAPTER II

REVIEW OF LITERATURE

In his book *The One Best System*, David Tyack (1974) used an illustration to picture the different roles and expectations of stakeholders in the public education system. This illustration, reproduced below as Figure 2, shows how each group of stakeholders had a different "picture" (or expectation) of what that system should look like; and each group had a different role to play in that system.

For this study, the illustration can be viewed on two levels. On a more macro level, it can be viewed as a depiction of science education reform and its stakeholders, each of whom has a different view as to what (reformed) science education should look like and each has a different role to play in the reform of science education in the United States.

On a more micro level, this illustration can be seen to depict the test development process. In this study, the process took place within a project evaluation; the project took place within a larger science education reform program; and the program took place within science education reform. Each group of stakeholders had their own expectation as to what a "science test" should look like; e.g., what its purpose and use should have been. Each group of stakeholders had a different role to play in this process. For

instance, the project group had a more direct role to play, and thus greater influence, in the creation of the science tests.

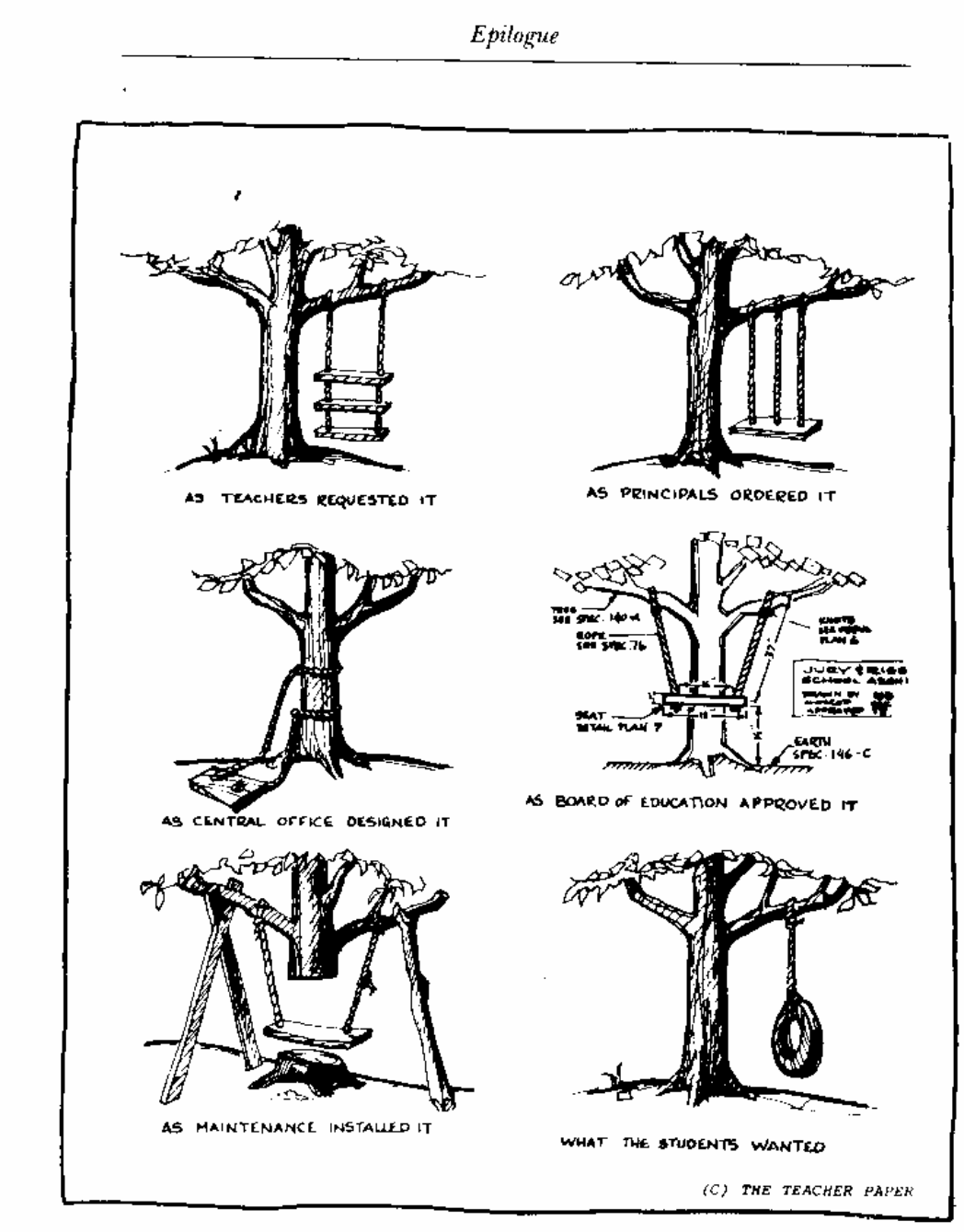


Figure 1. "What the Students Wanted." *The Teacher Paper* (n.d.), Portland, Oregon.

As stated in Chapter One, this dissertation focuses on the gap between theory (i.e., what one would have expected to happen) and practice (i.e., what actually happened) that arose in the process of developing a series of elementary grade science assessments for a funded NSF Math-Science Partnership project. The objective of this chapter is to present potential influences on the test development process within a project evaluation.

Four embedded rings (Figure 2) are used to represent not only the context in which this test development process took place but also potential sources of influence on this process. The outer ring, representing science education reform, constitutes the

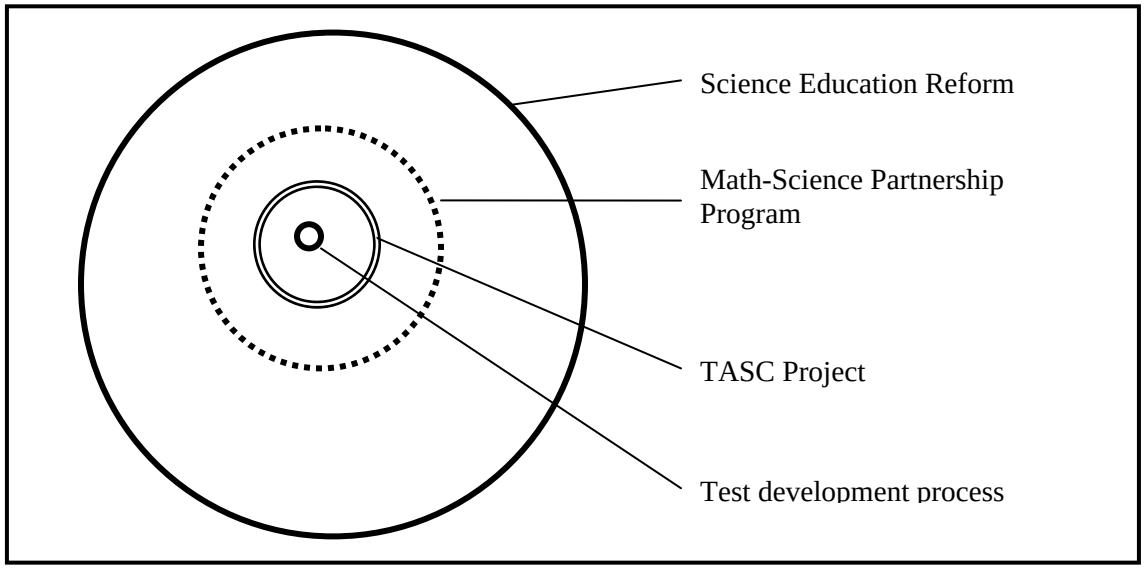


Figure 2. Context of test development process

"big picture" environment in which this study took place. Project 2061, launched in 1985 by the American Association for the Advancement of Science (AAAS), is an example of a long-term effort to reform science, mathematics, and technology education. Moving inward to the first ring—within science education reform—are many programs, for

example, the National Science Resource Center's Leadership Assistance for Science Education Reform (LASER) and—the program pertinent to this study—the National Science Foundation's Math-Science Partnership (MSP) program. Again moving inward to the next ring—and in the same way that there are many programs within science education reform, there are many projects within NSF's Math-Science Partnership program. The project pertinent to this study and represented in Figure 2 is the Teachers and Scientists Collaborating (TASC) project. Finally, the innermost ring in Figure 2 represents one task—the process of developing tests—to collect evidence concerning the effectiveness of the TASC project.

While there is more than one way to visualize this study, four embedded rings were selected to focus the discussion and to depict how each ring is nested within a larger ring. As stated previously, the inner rings are but one of many rings (i.e., programs, projects, tasks). However, only the ring of interest to this study was selected—that is, of the many science education reform programs, only the MSP program was selected; of the many MSP programs, only the TASC project was selected; and of the many project evaluation tasks, only the test development process was selected.

The discussion begins with the "big picture" context, or the most general (potential) influence on this test development process—current science education reform. Standards most reflective of current science education reform are the *National Science Education Standards (NSES).* The *Standards*, along with their influence on teacher science content knowledge and pedagogy, on assessment in science education and teachers' assessment literacy, and on North Carolina science education are presented.

The discussion then drills down to the next level of context, and potential influence, to one science education reform program (of many programs)—the National Science Foundation's Math-Science Partnership (MSP) program. Again drilling down, the discussion moves to one MSP project (of many projects)—the Teachers and Scientists Collaborating project—and then lastly, the discussion moves to one activity within the TASC project evaluation, i.e., the development of science tests.

The last section of the chapter presents the theoretical test development process and the standards recommended by the Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education associated with testing and measurement.

Current Science Education Reform

Even though current science education reform is the subject of this section, the discussion is directed toward the *National Science Education Standards*—six standards that outline what scientifically literate students should know, understand, and be able to do—because these *Standards* represent the direct influence of current science education reform on science education reform programs. Two standards—standards for science teaching and standards for assessment in science education—are singled out for additional discussion because they are directly pertinent to the factors that influenced the development of the science tests for this dissertation.

Social and economic challenges, as well as academic purposes, are motivating factors in current science education reform. Not only has knowledge of science and

technology become essential for average citizens as they make decisions about personal and social matters, such as health, natural resources, environment, and safety, but an expanding global economy demands a work force well-educated in science and technology. In response to these social and economic challenges, contemporary science education reform documents define what all students should know and be able to do in science in order to participate effectively in society. (Lee, 1998).

In 1983, the National Commission on Excellence in Education issued its report—*A Nation at Risk: The Imperative for Educational Reform*—that argued that academic standards had fallen in the U.S. demonstrated by low test scores of American youth, especially in math and science, and that this poor academic performance was the reason for its declining economic position in the world. (DeBoer, 2000). Beginning with *A Nation at Risk*, reports proclaiming the need to improve American education and providing numerous recommendations proliferated. Those recommendations associated with the quantity of science education—increasing required courses, school days, and the length of the school year—were implemented first because they were easiest. What remained were the more difficult aspects of educational quality and appropriateness: improving and coordinating curriculum, instruction, and assessments, and—especially critical—implementing those changes in the nation's classrooms. (Bybee, 1997).

By the late 1980s, there were more than 300 reports, all admonishing those within the educational system to change and consistently pointing out the specific need for reform in science education. Recommendations emphasized such issues as updated scientific and technological knowledge, application of learning theory and teaching

strategies, different approaches to achieving equity, and better preparation of students for the workplace. (Bybee, 1997).

In the 1980s, the general public began questioning the appropriate balance between science and technology and society. Society's perceptions of science and technology have important implications for science education policies, programs, and practices by highlighting a view of scientific literacy that requires more than an understanding of the concepts of traditional scientific disciplines. Citizens must be able to understand science in a social context, its interdependence with technology, and the nature and processes of both science and technology. These themes set the stage for a "general education" view of scientific literacy and established the perspective of the major policy statements of the 1980s and 1990s—*Science for All Americans* (American Association for the Advancement of Science, 1989), *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993), and *National Science Education Standards* (National Research Council, 1996). (Bybee, 1997).

In response to the numerous critical reports of American public education, the American Association for the Advancement of Science (AAAS)—the world's largest federation of scientific and engineering societies with hundreds of affiliate organizations and hundreds of thousands of individual members including scientists, engineers, science educators, policy makers, and interested citizens—put science literacy at the top of its priority list. The AAAS's goals included furthering the work of scientists; facilitating cooperation among them; fostering scientific freedom and responsibility; improving the effectiveness of science in the promotion of human welfare; advancing education in

science; and increasing public understanding and appreciation of the importance and promise of the methods of science in human progress. (AAAS, 1995, 1998).

In June 1985, the AAAS launched Project 2061, a long-term effort to reform science, mathematics, and technology education. The overall objective of Project 2061, which included 150 teachers and administrators in six school districts along with its advisory board—the National Council on Science and Technology Education, the American Association for the Advancement of Science, and numerous other team members and consultants, was to help shape the future of American education in which *all* students would become literate in science, mathematics, and technology by graduation from high school. (AAAS, 1995, 1993).

The first Project 2061 publication, *Science for All Americans (SFAA)*, was based on the premise that the scientifically literate person is one who is aware that science, mathematics, and technology are interdependent human enterprises with strengths and limitations; understands key concepts and principles of science; is familiar with the natural world and recognizes both its diversity and unity; and uses scientific knowledge and scientific ways of thinking for individual and social purposes. (AAAS, 1989). Project 2061's second publication, *Benchmarks for Science Literacy*, specified how students should progress toward science literacy. *Benchmarks* are statements of what *all* students should know or be able to do in science, mathematics, and technology by the end of grades 2, 5, 8, and 12. (AAAS, 1993).

National Science Education Standards

The *National Science Education Standards,* based in part on Project 2061's *Benchmarks for Science Literacy* and founded on exemplary practice and research, are designed to guide the U.S. toward a scientifically literate society. The *Standards* define scientific literacy as "the knowledge and understanding of scientific concepts and processes required for personal decision making, participation in civic and cultural affairs, and economic productivity." (NRC, 1996, p. 22) Specifically, the *Standards* outline what students need to know, understand, and be able to do to be considered scientifically literate at different grade levels. The *Standards* portray an educational system in which *all* students demonstrate high levels of performance, in which teachers make decisions essential for effective learning, in which interlocking communities of teachers and students focus on learning science (rather than learning *about* science), and in which supportive educational programs and systems nurture achievement. (NRC, 1996).

The *Standards*' development was guided by the following principles: (a) science is for all students; (b) learning science is an active process; (c) school science reflects the intellectual and cultural traditions that characterize the practice of contemporary science; and (d) improving science education is part of systemic education reform. (NRC, 1996).

The *Standards* are divided into the following six areas:

1 Standards for science teaching

2 Standards for professional development for teachers of science

3 Standards for assessment in science education

4 Standards for science content

5 Standards for science education programs

6 Standards for science education systems

Each set of standards will be presented and discussed in the following section.

1. Standards for Science Teaching

These standards are presented first to emphasize that effective teaching is at the center of science education. In addition to providing criteria for making judgments about progress toward the vision of science education as presented in the *Standards,* these standards also describe what teachers of science at all grade levels should understand and be able to do. The science teaching standards cover six areas: planning of inquiry-based science programs; actions taken to guide and facilitate student learning; assessments made of teaching and student learning; development of environments that enable students to learn science; creation of communities of science learners; and planning and development of the school science program. (NRC, 1996).

The standards for science teaching are based on five assumptions:

a. The vision of science education described by the *Standards* requires changes throughout the entire system.
b. What students learn is greatly influenced by how they are taught.
c. The actions of teachers are deeply influenced by their perceptions of science as an enterprise and as a subject to be taught and learned.
d. Student understanding is actively constructed through individual and social processes.

e. Actions of teachers are deeply influenced by their understanding of, and relationships with, students.

Because the *Standards* envision change occurring throughout the system, the teaching standards incorporate the following changes in emphases:

Less Emphasis On:	More Emphasis On:
o Treating all students alike and responding to the group as a whole	o Understanding and responding to individual student's interests, strengths, experiences, and needs
o Rigidly following curriculum	o Selecting and adapting curriculum
o Focusing on student acquisition of information	o Focusing on student understanding and use of scientific knowledge, ideas, and inquiry processes
o Presenting scientific knowledge through lecture, text, and demonstration	o Guiding students in active and extended scientific inquiry
o Asking for recitation of acquired knowledge	o Providing opportunities for scientific discussion and debate among students
o Testing students for factual information at the end of the unit or chapter	o Continuously assessing student understanding

Less Emphasis On:	More Emphasis On:
o Maintaining responsibility and authority	o Sharing responsibility for learning with students
o Supporting competition	o Supporting a classroom community with cooperation, shared responsibility, and respect
o Working alone	o Working with other teachers to enhance the science program

(NRC, 1996).

Teacher science content knowledge and pedagogy.

The *National Science Education Standards* portray a vision of science teaching and learning where students—helped to construct their own understanding of important science concepts—learn both the disciplinary content knowledge and how that knowledge is created. According to the *Standards*, students need to be engaged in genuine inquiries where they do not know the outcome beforehand; and at least some of the time, they need to have input in choosing the object of inquiry and designing the investigation. The assessment of students should be ongoing and used as much as practicable to monitor student progress and inform instructional decisions such as assigning grades. In standards-based instruction, the teacher functions as a facilitator of student learning rather than as a dispenser of information. Thus, for the *Standards* to impact student learning, they have to affect what happens in the science classroom and

that in turn depends in large part on teachers' knowledge, skills, and dispositions. (Horizon Research, 2003).

In 1977, the National Science Foundation initiated a major assessment of science and mathematics education throughout the United States. The first study, conducted in 1977, consisted of a comprehensive review of the literature, case studies of 11 districts throughout the U.S., and a national survey of teachers, principals, and district and state personnel. A second survey, conducted in 1985-86, surveyed teachers and principals to identify trends since 1977; and a third survey was conducted in 1993. (Weiss, et al., 2001).

The 2000 National Survey of Science and Mathematics Education—the fourth in this NSF-sponsored series of studies—was designed to provide the educational community with accurate and current information about science and mathematics education and trends in the following areas:

- Science and mathematics course offerings and enrollments;
- Availability of facilities and equipment;
- Instructional techniques;
- Textbook usage;
- Teacher background; and
- Needs for in-service education.

(Weiss, et al., 2001).

A stratified random sample of 1,800 schools in more than 1,200 school districts throughout the United States was selected to participate in the 2000 National Survey

with approximately 9,000 teachers selected for the survey from lists of mathematics and science teachers provided by school principals. Horizon Research, Inc. (Chapel Hill, NC), under the direction of Dr. Iris R. Weiss, conducted the survey. Westat, Inc. (Rockville, MD) was responsible for the data collection. In February 2000, survey questionnaires were mailed to individual teachers and department heads; data collection concluded in December 2000. (Weiss, et al., 2001).

One of the questions addressed by the 2000 National Survey was: How well prepared are science and mathematics teachers in terms of both content and pedagogy? Based on data collected for the 2000 National Survey, Weiss, et al. (2001) concluded that science and mathematics teachers, especially in the elementary and middle grades, do not have strong content preparation in their respective subjects. The authors found that whereas elementary teachers are usually assigned to teach science, mathematics, and other academic subjects to one group of students, the teachers did not feel equally qualified in each area. In fact, the authors found that of the elementary teachers surveyed:

- approximately 75 percent perceived themselves to be very well qualified to teach reading/language arts;
- approximately 60 percent perceived themselves to be very well qualified to teach mathematics, and
- approximately 25 percent perceived themselves to be very well qualified to teach science.

The authors opined that these results may be due to very few grade K-4 science and mathematics teachers having undergraduate majors in these fields, with the majority having majors in education.

In addition, evidence from the 1993 and 2000 National Surveys of Science and Mathematics Education suggested there had been no improvement in elementary teachers' preparedness to teach life science, earth science, or mathematics. (Smith, et al., 2002).

However, Weiss, et al. (2001) also found that even though fifth through eighth grade science and mathematics science teachers were more likely than their kindergarten through fourth grade colleagues to have undergraduate majors in science or mathematics, a majority had majors in education. Ninth through twelfth grade science and mathematics teachers, on the other hand, were found to be much more likely to have majored in their discipline than in education. The authors indicated that the number of semesters of college coursework completed by teachers revealed similar findings: elementary teachers had less extensive backgrounds than did their middle grade counterparts, who in turn had less science/mathematics coursework than their high school counterparts.

Weiss, et al. (2001) found that science teachers as a whole were much less likely to be familiar with the National Research Council's *National Science Education Standards* than mathematics teachers were with the National Council of Teachers of Mathematics *Standards.* In addition, they found that, in both subjects, teachers in the higher grades were more likely to be familiar with the respective *Standards* than teachers

in the lower grades; and that approximately 70 percent of the teachers familiar with the respective *Standards* agreed with its vision and indicated that they were implementing the recommendations of the *Standards* at least to a moderate extent.

Another area examined by Weiss, et al. (2001) was professional development. Here, they found that teachers indicated they do not have time during the school day to collaborate with their colleagues on issues of teaching science and mathematics. They also found that across subjects and grade ranges, teachers perceived their greatest need for their own professional development was learning how to use technology for instruction. Among K-8 science teachers, deepening their content knowledge ranked a close second. The authors stated that, by the teachers' own accounts, elementary science teachers were the most in need of professional development yet the least likely to participate in it. The authors also found that participation in professional development activities related to science and mathematics teaching was generally low, particularly among teachers in grades K-8 where less than 25 percent of the teachers had spent four or more days in professional development related to these subjects over the last three years.

In a report describing the status of elementary (grades K-5) school science instruction—based on the responses of 655 science teachers, 320 grade K-2 teachers, and 335 grade 3-5 teachers from the 2000 National Survey—Fulp (2002a) found that elementary school science teachers were lacking in content preparation, particularly in the physical sciences, and that almost 75 percent of the K-5 science teachers perceived a substantial need for professional development to deepen their own science content knowledge. The author indicated that the elementary school science teachers expressed a

need for help in using instructional technology and increasing their own content knowledge, but they spent very little time in professional development specific to science or science teaching, where they might receive such help.

In contrast, however, the elementary school science teachers reported a high degree of pedagogical preparedness, consistent with the high percentage of grade K-5 teachers of science who possess a degree in education. In general, these teachers reported feeling well prepared to implement more general pedagogical practices—i.e., listening and asking questions of their students and engaging their students in hands-on work and cooperative groups—than practices thought to be closely aligned with science standards—i.e., developing students' conceptual understanding of science, making connections between science and other disciplines, and leading students using investigative strategies. Teachers were less likely to indicate being well-prepared in the issue of technologies, particularly the use of computers for laboratory simulations and the use of the Internet for collaborative projects. (Fulp, 2002a).

In a separate report describing the status of middle school (grades 6-8) school science instruction—based on the responses of 529 middle school science teachers from the 2000 National Survey—Fulp (2002b) acknowledged that because the majority of middle school science classes are either general science or integrated science, teachers needed to possess a broad array of science content knowledge. However, she found that many middle school science teachers had gaps in their science content preparation; and thus it was not surprising to learn that relatively few middle school science teachers reported feeling well qualified to teach specific science concepts, with more than half

perceiving a substantial need for professional development to deepen their own science content knowledge.

Along with the elementary school science teachers, middle school science teachers reported a high degree of pedagogical preparedness, also consistent with the high percentage who possessed a degree in education. High percentages of middle school science teachers reported feeling well prepared to listen and ask questions of their students, to engage their students in hands-on work and cooperative groups, and to develop their students' conceptual understanding of science. Approximately one-third of the middle school science teachers reported being at least fairly familiar with the NRC *National Science Education Standards*, with over two-thirds of those agreeing with the *Standards'* vision and indicating that they were implementing the *Standards* in their classrooms. Similar to the elementary teachers of science, middle school science teachers were less likely to report being well prepared in the use of technologies, i.e., the use of computers for laboratory simulations and the use of the Internet for collaborative projects. (Fulp, 2002b).

Middle school science teachers reported spending very little time in professional development specific to science or science teaching. Approximately 25 percent of the middle school science teachers indicated they had not taken a course in science or the teaching of science since 1990. While middle school science teachers indicated the need for help in accommodating students with special needs, it appeared that little of the professional development in which they did participate focused on this area. (Fulp, 2002b).

In summary, the *NSES* teaching standards acknowledge that effective teaching is at the heart of science education, requiring teachers to continually expand their theoretical and practical knowledge about science, learning, and science teaching. (NRC, 1996). However, there appears to be a distinct mismatch between how teachers, particularly K-8 grade teachers of science, are prepared to teach science and how they are expected to teach science. In addition, teachers cannot assess what they themselves have not learned and have not taught. This, in turn, can be problematic where teachers are asked to write test questions on science content for which they have insufficient knowledge and experience.

2. Standards for Professional Development for Teachers of Science

These standards focus on four areas: the learning of science content through inquiry; the integration of knowledge about science with knowledge about learning, pedagogy, and students; the development of the understanding and ability for lifelong learning; and the coherence and integration of professional development programs. (NRC, 1996).

These two foregoing sets of standards present a view of science teaching that is based on the conviction that scientific inquiry is central to science and science learning. (NRC, 1996.)

3. Standards for Assessment in Science Education

The assessment standards provide criteria against which to judge the quality of assessment practices, including classroom-based and externally designed assessments. These standards cover five areas: the consistency of assessments with the decisions they

are designed to inform; the assessment of both achievement and opportunity to learn science; the match between the technical quality of the data collected and the consequences of the actions taken on the basis of those data; the fairness of assessment practices; and the soundness of inferences made from assessments about student achievement and opportunity to learn. (NRC, 1996).

Because the current research project's central focus is the test development process, the assessments standards of the *National Science Education Standards* are discussed more fully. The assessment standards provide criteria to judge progress toward the science education vision of scientific literacy for all, describing the quality of assessment practices used by teachers and state and federal agencies to measure student achievement and the opportunity provided students to learn science. The standards identify essential characteristics of exemplary assessment practices and thus serve as guides for developing assessment tasks, practices, and policies. The standards can be applied equally to the assessment of students, teachers, and programs; to summative and formative assessment practices; and to classroom assessments as well as large-scale, external assessments. (NRC, 1996).

The assessments standards include five substandards:

- *Assessment Standard A: Assessments must be consistent with the decisions they are designed to inform.*
    - *Assessments are deliberately designed.*

      Evidence of such deliberate design may be found in written plans for assessments that contain:

- statements about the purposes that the assessment will serve;
- descriptions of the substance and technical quality of the data to be collected;
- specifications of the number of students or schools from which data will be obtained;
- descriptions of the data-collection method;
- descriptions of the method of data interpretation; and
- descriptions of the decisions to be made, including who will make the decisions and by what procedures.

- *Assessments have explicitly stated purposes*.

  Because conducting assessments is such a resource-intensive activity, it should not be undertaken unless there is assurance that the subsequent decisions and actions will increase the scientific literacy of the students—an assurance that can be made only if the purpose of the assessment is clear.

- *The relationship between the decisions and the data is clear*.

- *Assessments procedures are internally consistent*.

- *Assessment Standard B: Achievement and opportunity to learn science must be assessed*.

  - *Achievement data collected focus on the science content that is most important for students to learn*.

    The science content standards portray the outcomes of science education, including:

- the ability to inquire;
- knowing and understanding scientific facts, concepts, principles, laws, and theories;
- the ability to reason scientifically;
- the ability to use science to make personal decisions and to take positions on societal issues; and
- the ability to communicate effectively about science.

This assessment standard makes clear the complexity of the content standards while addressing the importance of collecting data on all aspects of student science achievement. Assessments need to probe the extent and organization of a student's knowledge, including reasoning and utilization of such knowledge.

- *Opportunity-to-learn data collected focus on the most powerful indicators.* Some of those indicators, at the classroom level, mentioned in this standard include teachers' professional knowledge, i.e., content knowledge, pedagogical knowledge, and understanding of students; the extent to which content, teaching, professional development, and assessment are coordinated; the time available for teachers to teach and students to learn science; the availability of resources for student inquiry; and the quality of educational materials available.
- *Equal attention must be given to the assessment of opportunity to learn and to the assessment of student achievement.*

- *Assessment Standard C: The technical quality of the data collected is well matched to the decisions and actions taken on the basis of their interpretation.*

  This standard addresses the degree to which the data collected warrant the decisions and actions that will be based on them.

  - *The feature that is claimed to be measured is actually measured.*
  - *Assessment tasks are authentic.*

    Authentic assessment tasks are defined as tasks that are similar in form to tasks in which students will engage in their lives outside the classroom or are similar to the activities of scientists.
  - *An individual student's performance is similar on two or more tasks that claim to measure the same aspect of student achievement.*
  - *Students have adequate opportunity to demonstrate their achievements.*

    According to this standard, assessment tasks must be developmentally appropriate, must be set in contexts that are familiar to the students, must not require reading skills or vocabulary that are inappropriate to the students' grade level, and must be as free from bias as possible.
  - *Assessment tasks and methods of presenting them provide data that are sufficiently stable to lead to the same decisions if used at different times.*

    One aspect of reliability, this is particularly important for large-scale assessments, where changes in performance of groups are of interest. It is only with stable measures that valid inferences can be made about changes in group performance.

Data-collection methods can take different forms, each with distinct advantages and disadvantages. To serve the intended purpose of the assessment, the choice of assessment form should be consistent with what one wants to measure and to infer. Thus, it is imperative that the data and their method of collection yield information with confidence levels consistent with the consequences of its use.

- *Assessment Standard D: Assessment practices must be fair.*

  Because one of the premises of the *Standards* is that all students should have access to quality science education and should be expected to achieve scientific literacy as defined by the content standards, it follows that the processes used to assess student achievement must be fair to all students.

  - *Assessment tasks must be reviewed for the use of stereotypes, for assumptions that reflect the perspectives or experiences of a particular group, for language that might be offensive to a particular group, and for other features that might distract students from the intended task.*

    Planners and implementers of science assessments must pay deliberate attention to issues of fairness that should be reflected in the procedures used to develop the assessment tasks, in the content and language of the assessment tasks, in the processes by which students are assessed, and in the analyses of assessment results.

  - *Large-scale assessments must use statistical techniques to identify potential bias among subgroups.*

- *Assessment tasks must be appropriately modified to accommodate the needs of students with physical disabilities, learning disabilities, or limited English proficiency.*
- *Assessment tasks must be set in a variety of contexts, be engaging to students with different interests and experiences, and must not assume the perspective or experience of a particular gender, racial, or ethnic group.*

- *Assessment Standard E: The inferences made from assessments about student achievement and opportunity to learn must be sound.*
  - *When making inferences from assessments about student achievement and opportunity to learn science, explicit reference needs to be made to the assumptions on which inferences are based.*

    Even when assessments are well planned, yielding high quality data, the interpretations of the empirical evidence can result in quite different conclusions. Making inferences involves examining empirical data while looking through the lenses of theory, personal beliefs and personal experience. Because individuals are not always aware of the assumptions they make, confidence in the validity of inferences requires explicit reference to the assumptions on which those inferences are based. The level of confidence in conclusions is increased when those conducting assessments have been well trained in the process of making inferences from educational assessment data.

(NRC, 1996).

Overall, the assessment standards of the *National Science Education Standards* include the following changes in emphases:

Less Emphasis On:	More Emphasis On:
o Assessing what is easily measured	o Assessing what is most highly valued
o Assessing discrete knowledge	o Assessing rich, well-structured knowledge
o Assessing scientific knowledge	o Assessing scientific understanding and reasoning
o Assessing to learn what students do not know	o Assessing to learn what students do understand
o Assessing only achievement	o Assessing achievement and opportunity to learn
o End-of-term assessments by teachers	o Students engaged in ongoing assessment of their work and that of others
o Development of external assessments by measurement experts alone	o Teachers involved in the development of external assessments

Assessment literacy.

Popham (1999) stated that while teachers like to teach, they rarely like to test. However, he asserted that effective testing enhances a teacher's instructional effectiveness and thus teachers who can test well will be better teachers. Popham (1999) offered four "traditional" reasons why teachers should know about assessment: Testing

enables teachers to diagnose their students' strengths and weaknesses, to monitor their students' progress, to assign students' grades, and to determine their own instructional effectiveness. Additionally, he cited three "contemporary" reasons why teachers should know about assessment:

- Test results determine public perceptions of educational effectiveness.
- Students' assessment performances increasingly are seen as part of the teacher evaluation process.
- Assessment instruments—as clarifiers of instructional intentions—can improve instructional quality.

These reasons supporting the importance of knowing about assessment are linked to decisions. As stated previously, because educational decision-making is at the center of education testing, the more psychometrically sound an assessment instrument, the more confident one can be in the decisions based on that assessment's results.

From a January 1998 survey of state teacher licensing standards, Stiggins (1999) found that only 25 of 50 states require that teachers either meet specific assessment competence standards or at least complete assessment course work during their preparation. Specifically, 15 states with teacher certification standards require competence in assessment: CO, CT, DE, FL, HI, IN, NY, OH, OK, OR, TX, UT, VT, VA, WA; 10 states explicitly require assessment course work during training: AL, AK, AZ, CA, IA, MT, ND, TX, WI, WY; and 25 states hold no expectation of competence in assessment: AR, GA, ID, IL, KS, KY, LA, ME, MD, MA, MI, MN, MS, MO, NE, NV, NH, NJ, NM, NC, PA, RI, SC, SD, WV. Stiggins (1999) pointed out that although this

count is up sharply from previous surveys in 1983, 1988, and 1991, much remains to be done.

Professional associations of educators have rallied around the need for assessment literacy in the classroom and school building. In fact, almost every set of standards of teacher competence developed recently, including those developed by the National Education Association (NEA), the American Federation of Teachers (AFT), the Council of Chief State School Officers (CCSSO), the National Council for Accreditation of Teacher Education (NCATE), and the National Board of Professional Teacher Standards (NBPTS), holds the expectation that teachers will be competent in assessment. (Stiggins, 1999).

However, even at this time with increased emphasis on testing and assessment, many colleges of education and state education agencies do not require preservice teachers to complete specific coursework in classroom assessment (Campbell et al., 2002). Plake (1993) found that many inservice teachers reported they were not well prepared to assess student learning and that this lack of adequate preparation was largely due to inadequate preservice teacher training in the area of educational measurement. Brookhart (2001), citing literature that called for an increase in emphasis in teacher preparation programs on classroom assessment and a decrease in emphasis on large-scale testing, summarized the research on teachers' assessment practices by stating that teachers apparently do better at classroom applications than at interpreting standardized tests and that they lack expertise at test construction.

What is “assessment literacy”?

Assessment literacy has been defined by some as knowing how to assess what students know and can do, including interpreting results from such assessments and applying the results to improve student learning and program effectiveness. (Webb, 2002). Others have defined it as possessing knowledge about the basic principles of sound assessment practice, including terminology, the development and use of assessment methodologies and techniques, administration, analysis, and familiarity with standards of quality in assessment (SCASS, 2003).

Stiggins (1995), rather than formally defining assessment literacy, described "assessment literates" as those who know the difference between sound and unsound assessment and "are not intimidated by the sometimes mysterious and always daunting technical world of assessment” (p. 240). He further stated that assessment-literate educators (i.e., teachers, administrators, superintendents) enter the realm of assessment knowing what they are assessing, why they are doing it, how best to assess the skill/knowledge of interest, how to generate good examples of student performance, what can potentially go wrong with the assessment, and how to prevent that from happening. In addition, they are aware of the potential negative consequences of poor, inaccurate assessment. (Stiggins, 1995).

Assessment literacy is a key component of *The Standards for Teacher Competence in the Educational Assessment of Students* (American Federation of Teachers, National Council on Measurement in Education, and National Education Association, 1990). These standards—a joint effort between the American Federation of

Teachers, National Council on Measurement in Education, and National Education Association—were initiated in 1987 to address the problem of inadequate assessment training for teachers (AFT et al., 1990). Even though assessments play a pivotal role in the No Child Left Behind Act of 2001, these standards remain unchanged.

The 1990 *Standards* define assessment as "the process of obtaining information that is used to make educational decisions about students, to give feedback to the student about his or her progress, strengths, and weaknesses, to judge instructional effectiveness and curricular adequacy, and to inform policy" (AFT et al., 1990). The *Standards* consist of the following seven principles:

- Standard 1: Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
- Standard 2: Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
- Standard 3: The teacher should be skilled in administering, scoring, and interpreting the results of both externally produced and teacher-produced assessment methods.
- Standard 4: Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
- Standard 5: Teachers should be skilled in developing valid pupil grading procedures that use pupil assessments.

- Standard 6: Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
- Standard 7: Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

These standards, which acknowledge the importance of teacher education and professional development in the area of classroom assessment, apply to teachers' development and use of classroom assessments based on the instructional goals and objectives that form the basis for classroom instruction. In addition, Standards 3, 4, 6, and 7 apply to large-scale assessment, including administering, interpreting, and communicating assessment results, using information for decision making, and recognizing unethical practices. (Brookhart, 2001).

Research on assessment literacy and the *Standards.*

Although numerous studies have been conducted over the past decade addressing one or more of the seven *Standards* (Brookhart, 2001), a few studies (Plake, 1993; Zhang & Burry-Stock, 1995) addressed all teacher competencies as specified by the *Standards* for inservice teachers. Campbell, et al. (2002) attempted to apply the *Standards* to groups of undergraduate preservice teachers. Mertler (2004) applied the *Standards* to secondary preservice and inservice teachers.

By using the *Standards* as a blueprint for the development of a survey instrument, Plake (1993) created an instrument—the Teacher Assessment Literacy Questionnaire—made up of five application-type items per Standard. Once finalized, the TALQ was administered to a representative sample of teachers from 98 districts in 48 states resulting

in a total usable sample of 555 surveys. Answering an average of 23 out of 35 questions correctly, teachers scored highest on Standard 3 (Administering, Scoring, and Interpreting the Results of Assessments) with a mean score of 3.96 out of 5.00 and lowest on Standard 6 (Communicating Assessment Results) with a mean score of 2.70 out of 5.00. (Plake, 1993).

Zhang and Burry-Stock (1995) used the Assessment Practices Inventory (API; Zhang & Burry-Stock, 1994)—also developed based on the seven *Standards*—to investigate inservice teachers' assessment competency as a function of measurement training and years of teaching. Data from 311 inservice (elementary, middle school/junior high, and high school) teachers from two local school districts in Alabama were collected and analyzed. Results suggested that among teachers who had taught for four or more years, those who had received measurement training were more skilled in interpreting standardized test results, conducting classroom statistics , and using assessment results in decision making than those who had not received any measurement training. Additionally, these authors found that those who had received measurement training were more skilled in using performance assessment and informal observation than those who had not received any measurement training.

In a subsequent study, Zhang and Burry-Stock (2003)—again using the Assessment Practices Inventory—investigated teachers' assessment practices across teaching levels and content areas as well as teachers' self-perceived assessment skills as a function of teaching experience and measurement training. From their analyses of 267 completed APIs, Zhang et al. (2003) found that as grade level increased, teachers relied

more on objective tests in classroom assessment and showed an increased concern for assessment quality. The researchers also found that across content areas, teachers' involvement in assessment activities reflected the nature and importance of the subjects they taught. Lastly, Zhang, et al. (2003) found that regardless of teaching experience, teachers with measurement training reported a higher level of self-perceived assessment skills in using performance measures; in standardized testing, test revision, and instructional improvement; as well as in communicating assessment results than those without measurement training.

Zhang, et al.'s finding that knowledge in measurement and testing had a significant impact on teachers' self-perceived assessment skills regardless of their teaching experience confirmed teachers' beliefs that university coursework contributed to their knowledge of testing and measurement (see also Gullikson, 1984; Wise, Lukin, & Roos, 1991). Zhang, et al. (2003) concluded that this finding suggests that measurement training may compensate for novices' lack of experience in the classroom and that the results from their study provide evidence for the value of university coursework in tests and measurement.

Renaming the Teacher Assessment Literacy Questionnaire as the Assessment Literacy Inventory (ALI), Campbell, et al. (2002) conducted a similar study with undergraduate preservice teachers. The ALI was administered to 220 undergraduate students following a course in tests and measurement. The course included such topics as creating and critiquing various methods of assessment, discussing ethical considerations related to assessment, interpreting and communicating both classroom and standardized

assessment results, and discussing and evaluating psychometric qualities (i.e., validity and reliability) of assessments. The preservice teachers scored highest on Standard 1 (Choosing Appropriate Assessment Methods) and lowest on Standard 6 (Communicating Assessment Results). (Campbell, et al., 2002).

In fall 2002, Mertler (2004) surveyed 67 undergraduate students (science and social studies) at a midwestern university and 101 teachers, representing nearly every district and school in a three-county area surrounding the same institution. Mertler surveyed both groups using the *Classroom Assessment Literacy Inventory* (CALI), that consisted of the same 35 content-based items as the *Teacher Assessment Literacy Questionnaire* (Plake, 1993) with a limited amount of rewording and the addition of seven demographic items. Mertler found that, for the inservice teachers, the highest mean performance was associated with Standard 3 (Administering, Scoring, and Interpreting Assessment Results) and lowest with Standard 5 (Developing Valid Grading Procedures). For the preservice teachers, Mertler found the highest mean performance was associated with Standard 1 (Choosing Appropriate Assessment Methods) and lowest with Standard 5. Statistically significant differences ($p<.05$) between the two groups were found for Standards 1, 2, 3, 4, 7, and total score.

Mertler (2004) concluded that preservice training of secondary teachers in the concepts and techniques of classroom assessment is critical and should be enhanced through thoughtful examination and research into the knowledge and skills that secondary teachers need to possess once they assume the responsibilities of their own classroom students. In addition, Mertler (2004) recommended that ongoing training on

various topics related to classroom assessment be an essential component of any district's program of professional development for its secondary teachers.

4. Standards for Science Content

The content standards provide expectations for the development of student understanding and ability over the course of K-12 education. They outline what students should know, understand, and be able to do. These standards are divided into eight categories: unifying concepts and processes in science; science as inquiry; physical science; life science; earth and space science; science and technology; science in personal and social perspective; and history and nature of science. (NRC, 1996).

5. Standards for Science Education Programs

The program standards provide criteria for judging the quality of school and district science programs. These standards focus on six areas: the consistency of the science program with the other standards and across grade levels; the inclusion of all content standards in a variety of curricula that are developmentally appropriate, interesting, relevant to student's lives, organized around inquiry, and connected with other school subjects; the coordination of the science program with mathematics education; the provision of appropriate and sufficient resources to all students; the provision of equitable opportunities for all students to learn the standards; and the development of communities that encourage, support, and sustain teachers. (NRC, 1996).

The program standards make clear that assessment policies and practices should be aligned with the goals, student expectations, and curriculum frameworks and that the alignment of assessment with curriculum and teaching is one of the most critical pieces of

science education reform. In fact, the standards indicate that reform is undermined if the assessment system at the school and district levels does not reflect the *Standards* and measure what is valued. (NRC, 1996).

6. Standards for Science Education Systems

The system standards establish criteria for judging the performance of the overall science education system. These standards cover seven areas: the congruency of policies that influence science education with the teaching, professional development, assessment, content and program standards; the coordination of science education policies within and across agencies, institutions, and organizations; the continuity of science education policies over time; the provision of resources to support science education policies; the equity embodied in science education policies; the possible unanticipated effects of policies on science education; and the responsibility of individuals to achieve the new vision of science education portrayed in the standards. (NRC, 1996).

Influence on North Carolina Science Education

The science component of the North Carolina Standard Course of Study (SCS) was created by establishing competency goals and objectives for teaching and learning science in all grades. In addition to containing the concepts and theories, strands, skills, and processes on which all science instruction should be based, the 1999 revision of the SCS defined and illustrated the connections between the *National Science Education Standards*, the *Benchmarks for Scientific Literacy*, and the state standards. The SCS was revised in 2004 to better reflect the *National Science Education Standards* along with the 1996 National Assessment of Educational Progress (NAEP) science framework and

assessment. (North Carolina DPI, 2004). The 1996 NAEP framework was organized along a content dimension (earth, physical, and life sciences) and a cognitive dimension (conceptual understanding, scientific investigation, and practical reasoning). Each question in the 1996 and 2000 NAEP science assessments was categorized by its content and cognitive domains. (O'Sullivan, Lauko, Grigg, Qian, and Zhang, 2003).

In congruence with the *National Science Education Standards*, the overall goal of the 2004 North Carolina Standard Course of Study for *all* North Carolina students is to achieve scientific literacy as defined by those *Standards*, that is, the ability to: find or determine answers to questions derived from every day experiences; describe, explain, and predict natural phenomena; understand articles about science; engage in non-technical conversation about the validity of conclusions; identify scientific issues underlying national and local decisions; and pose explanations based on evidence derived from one's own work. (North Carolina DPI, 2004).

North Carolina Grade Level Science Competencies

In third grade, students focus on the study of systems as their unit of investigation, learning that a system is an interrelated group of objects or components that form a functioning unit. Because of the complexity of the natural world and human-designed systems, the third grade science program allows students to identify small components of a system for in-depth investigation. (NC DPI, 2004).

Building on their third grade science instruction, fifth grade students focus on using evidence, models, and reasoning to form scientific explanations. Evidence, consisting of observations and data on which scientific explanations are based, is used by

students to predict changes in natural and human-designed systems. Models—tentative schemes or structures constructed to represent real objects or processes—help students understand how things work. Explanations incorporate prior knowledge and new evidence from observations, experiments, or models into consistent, logical statements, and thus enable students to more precisely understand scientific concepts and processes. (NC DPI, 2004).

Eighth grade science provides investigational opportunities—integrated with mathematics, technology, and social science—help students learn about themselves and their world and how to communicate that learning to others. Thus, designing technological solutions and pondering benefits and risks should be an integral part of the middle school science experience. (NC DPI, 2004).

Inquiry—the use of the processes of science, scientific knowledge, and attitudes to reason and to think critically—is the manner by which students should learn science. The *NSES* state that inquiry supports a learner in constructing an understanding of scientific concepts, learning how to learn, becoming an independent and lifelong learner, and further developing habits of mind associated with science. Participating in inquiry-based science education, students should be able to ask questions, use their questions to plan and conduct a scientific investigation, use appropriate science tools and science techniques, evaluate evidence and use it logically to construct several alternative explanations, and communicate (defend) their conclusions scientifically. (NRC, 1996).

Because this dissertation focuses on the gap between theory and practice in the test development process within a project evaluation, it is necessary to understand the

context in which this process took place in order to consider potential influences on the process. This section highlighted one source of potential influence, that is, national science standards that are reflective of science education reform.

As stated previously, there are many programs within science education reform. The program of interest in this study is the National Science Foundation's Math-Science Partnership program that, in turn, directly influences the many projects within its purview. Thus, the National Science Foundation, through its Math-Science Partnership program, is highlighted in the next section as another potential source of influence on the test development process within a (MSP) project evaluation.

National Science Foundation's Math-Science Partnership Program

Established by Congress in 1950, the National Science Foundation (NSF) is the only federal agency dedicated to the support of education and fundamental research in all scientific and engineering disciplines. NSF's mission is to ensure that the United States maintains leadership in scientific discovery and the development of new technologies. (Retrieved 3/28/06 from http://www.nsf.gov/about/history/ .)

The National Science Foundation is divided into seven directorates, each headed by an assistant director and further subdivided into divisions, that support science and engineering research and education along with four offices that support research and researchers. NSF's Math-Science Partnership Program falls under the purview of the Directorate for Educational and Human Resources (EHR). The primary goals of the Directorate for Education and Human Resources include:

1. Preparation of the next generation of STEM professionals along with attracting and retaining more Americans to STEM careers.
2. Development of a robust research community that can conduct rigorous research and evaluation that will support excellence in STEM education and that integrates research and education.
3. Increasing the technological, scientific, and quantitative literacy of all Americans so that they can exercise responsible citizenship and live productive lives in an increasingly technological society.
4. Broadening participation (individuals, geographic regions, types of institutions, STEM disciplines) and closing achievement gaps in all STEM fields.

(Retrieved 3/28/06 from http://www.nsf.giv/her/about.jsp.)

NSF defines a program as "a coordinated approach to exploring a specific area related to NSF's mission of strengthening science, mathematics, and technology" and a project as "a particular investigative or development activity funded by that program." Thus, a program is made up of a collection of projects that seek to meet a defined set of goals and objectives. (NSF, 2002b).

In 2002, NSF initiated its Math and Science Partnership (MSP) program in response to the President's vision, legislated in the No Child Left Behind Act of 2001, to strengthen and reform prekindergarten through twelfth grade (preK-12) education. The MSP program, a collaboration between NSF and the U.S. Department of Education, seeks to improve student outcomes in high-quality mathematics and science by all students at all preK-12 grade levels. (NSF, 2002). In serving students and educators, the MSP

program supports the development, implementation, and sustainability of exemplary partnerships that:

- enhance schools' capacity to provide challenging curricula for all students and encourage more students to succeed in advanced courses in mathematics and the sciences;
- increase the number, quality, and diversity of mathematics and science teachers, particularly in underserved areas;
- engage and support scientists, mathematicians, and engineers at local universities and local industries to work with K-12 educators and students;
- contribute to a greater understanding of how students effectively learn mathematics and science and how teacher preparation and professional development can be improved; and
- promote institutional and organizational change in education systems—from kindergarten through graduate school—to sustain partnerships' promising practices and policies.

The MSP program is made up of four components:

- *Comprehensive Partnerships* implement change across the K-12 continuum in mathematics and/or science.
- *Targeted Partnerships* focus on improved student achievement in a narrower grade range or disciplinary focus in mathematics and/or science.
- *Institute Partnerships* develop mathematics and science teachers as school- and district-based intellectual leaders and master teachers.

- *Research, Evaluation, and Technical Assistance (RETA)* activities assist partnership awardees in the implementation and evaluation of their work.

(NSF, 2005).

In presenting the context within which this study takes place (Figure 2, page 15), two potential sources of influence on the test development process within a project evaluation have been presented and discussed—that is, science education reform (the outermost ring in Figure 2) as reflected in the *National Science Education Standards* and the National Science Foundation's Math-Science Partnership program (moving inward, the next ring), one of many science education reform programs.

As stated previously, NSF's Math-Science Partnership program includes many projects. The one of pertinence to this study, and the next potential source of influence on the test development process in this study, is the Teachers and Scientists Collaborating (TASC) project, the topic of the next section.

Teachers and Scientists Collaborating Project

The TASC project, a targeted MSP partnership, is a collaboration between the Pratt School of Engineering at Duke University and the Center for Inquiry-Based Learning (CIBL), an independent non-profit organization. Based upon information gleaned from Duke University's website and from working directly with TASC, Figure 3 illustrates the contractual connections between TASC, CIBL, and Duke University. As Figure 3 indicates, Duke University (the fiscal agent for the TASC project) has a collaborative relationship with CIBL, an advisory coalition of scientists

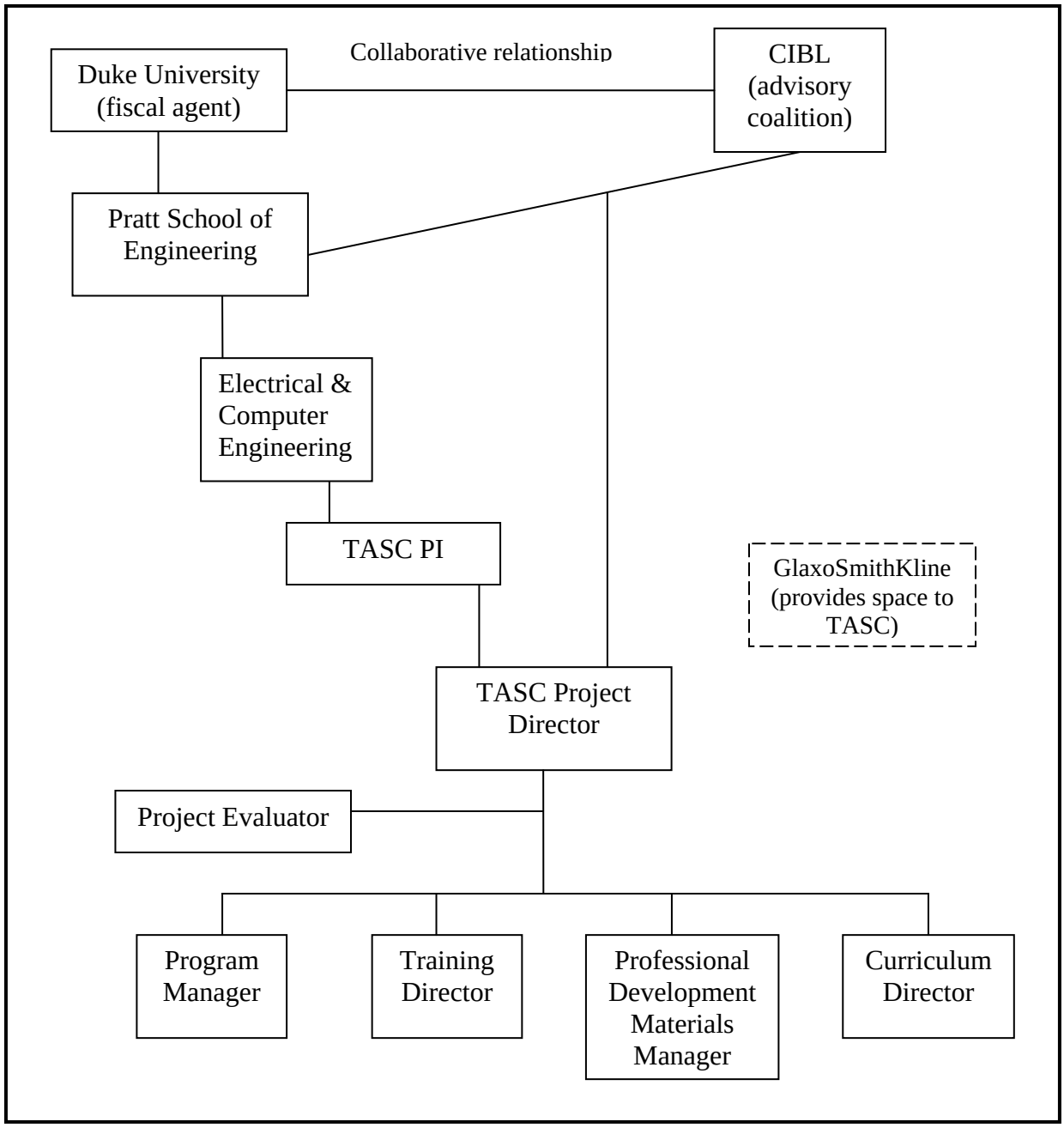


Figure 3. TASC's organization chart

working with Duke University's Pratt School of Engineering. Within the Pratt School of Engineering is the Electrical and Computer Engineering department in which the TASC Principal Investigator works. The TASC Project Director reports to the TASC Principal

Investigator and interacts with CIBL and Pratt's scientists. Parties reporting to the TASC Project Director are the Project Evaluator, the Program Manager, the Training Director, the Professional Development Materials Manager, and the Curriculum Director. GlaxoSmithKline provides office, training, and warehouse space to TASC.

Focusing on kindergarten through eighth grade science education, TASC assists science educators in shifting toward inquiry-based science teaching by providing curriculum units from a selection of NSF-approved, inquiry-based curricula; in-service professional development; and support from Duke University scientists. Specifically, TASC provides:

1. two days' intensive professional development in the use of selected (i.e., NSF-approved) curriculum units, background information, and inquiry-based teaching, including payment for substitute teachers (for TASC Partners) needed to fill in for teachers in training;
2. supply and refurbishment of NSF-supported curriculum units, matched to the North Carolina Standard Course of Study (NC SCS), that meet needs in North Carolina K-8 science classrooms; and
3. support from Duke University scientists trained in helping teachers use specific curriculum units. Teachers and scientists collaborate electronically and in person.

(Retrieved May 9, 2005, from http://tasc.pratt.duke.edu/ about.overview.php).

The professional development offered by TASC includes training on multiple curriculum units throughout the school year. (Appendix A provides the TASC training schedule for the 2005-2006 school year.) After teachers have registered for, and

attended, the TASC training on a particular curriculum unit, TASC sends the curriculum unit kits to the teachers' classrooms.

The science unit kits used by TASC are either K-8 FOSS (Full Option Science System), BSCS Science T.R.A.C.S. (Teaching Relevant Activities for Concepts and Skills), or SEPUP (Science Education for Public Understanding Program) curricula, all of which have been developed with support from the National Science Foundation. Each of the curriculum units purport to be aligned with the *National Science Education Standards*, the *Benchmarks*, and the science competency goals of the North Carolina Standard Course of Study. The science unit kits, that remain in the elementary teachers' classrooms for 9 weeks and in the middle school teachers' classrooms for 14 weeks, contain all the materials necessary for teachers to create learning environments in which students take the role as scientists. That is, students take the initiative to observe and question phenomena; they pose explanations of what they observe; they develop theories and conduct tests to support or refute them; they collect and analyze data; they draw conclusions based upon experimental data; they design and build models. In summary, through actively engaging students in inquiry-oriented experiences, teachers facilitate students' learning of science concepts and facts as well as the processes involved in establishing those concepts and facts. (Retrieved May 9, 2005, from http://tasc.pratt.duke.edu/ about.overview.php).

Each TASC training cycle serves approximately 20 teachers per grade level. For each unit, teachers receive training during the school year in two six-hour days, separated by three weeks. TASC training on each unit repeats annually. TASC projected that

training on new units for each grade level would be added annually through the project's duration so that training would be available for units covering most of the K-8 science competency goals in the NC Standard Course of Study. Participating teachers must receive TASC training on a curriculum unit before they may use it in their classrooms. In addition, TASC offers advanced training on inquiry-based teaching and in-depth content related to curriculum. (Retrieved May 9, 2005, from http://tasc.pratt.duke.edu/about.overview.php).

One of TASC's anticipated outcomes is improvement in students' and teachers' science content and process knowledge. TASC contracted for the development of science tests in order to determine whether its teacher training on the curriculum units improved teachers' science content and process knowledge and whether its training, and subsequent implementation in the classroom, improved students' science content and process knowledge. TASC expected to use results from the tests to help them evaluate the effectiveness of the teacher professional development being presented by TASC to science educators from the four North Carolina school districts (Alamance/Burlington Schools, Orange County Schools, Iredell/Statesville Schools, and Harnett County Schools) that initially participated in this project.

To examine the gap between theory and practice in the test development process within a project evaluation, this chapter presented potential sources of influence on this process. It began with the most general "ring" of influence—science education reform as reflected in the *National Science Education Standards*—and then moved to a more specific "ring" of influence—the National Science Foundation's Math-Science

Partnership program—and then finally to a more specific "ring" of influence within NSF's MSP program—the Teachers and Scientists Partnership project.

The next two sections—Project Evaluation and Test Development Process—present the final sources of influence on the test development process. The Project Evaluation section begins with a general, and brief, discussion of evaluation; then moves to an overview of NSF's expectations for the evaluations of projects within its various programs; and lastly turns to TASC's evaluation plan to NSF. The Test Development Process section presents the theoretical process by which tests are constructed and thus the most significant, from a test developer's perspective, influence on the test development process.

Project Evaluation

The Joint Committee on Standards for Educational Evaluation (1994) broadly defines evaluation as a "systematic investigation of the worth or merit of an object." Central to this definition is the notion of using evaluation for a purpose. That is, evaluations should be conducted for action-related reasons, and the information provided should facilitate a course of action. (NSF, 2002b).

Evaluations are undertaken for various reasons, such as, helping management improve a program; gaining knowledge about a program's efforts; providing input to decisions about the program's funding, structure, or administration; or responding to political pressures. Generally, an evaluation intended to provide information for guiding program improvement is called a formative evaluation; and an evaluation intended to

render a summary judgment on a program's performance is called a summative evaluation. (Rossi, et al., 2004).

NSF distinguishes between program evaluation and project evaluation. A program evaluation determines the value of the collection of projects making up the program (e.g., Math-Science Partnership program); and looking across projects, NSF examines the utility of the activities and strategies employed. A project evaluation, in contrast, collects information about the progress and outcomes of an individual project (e.g., TASC project). That is, information is collected to help determine whether the project is proceeding as planned and whether it is meeting its stated program goals and project objectives according to the proposed timeline. (NSF, 2002b).

The purposes of a project evaluation are to help project directors and principal investigators monitor and refine their own work and to provide information of value to NSF and to the field. Project evaluation generally examines two types of issues: implementation and outcomes. The principal investigator, typically working with an evaluator, is responsible for ensuring that a project evaluation is carried out. (Retrieved 3/6/06 from http://www.nsf.gov/her/rec/compareeval.jsp.)

Figure 4 illustrates the types of evaluation. Formative evaluation, which begins during project development and continues throughout the life of the project, assesses ongoing project activities and provides information to monitor and improve the project. Formative evaluation has two components: implementation evaluation, which assesses whether the project is being conducted as planned, and progress evaluation, which

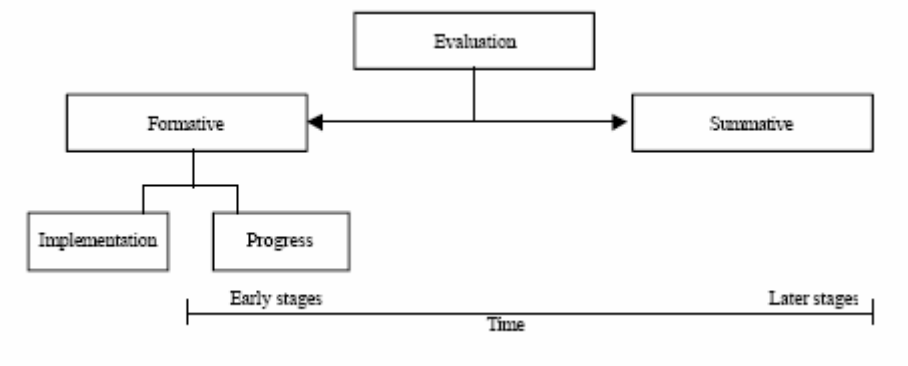


Figure 4. Types of evaluation (NSF, 2002b).

assesses progress in meeting the goals of the program and the project. Implementation evaluation collects information to verify that the program and its components are actually operating and whether they are operating according to the proposed plan or description. Progress evaluation collects information to ascertain what the impact of the activities and strategies is on participants, curriculum, or institutions at various stages of the intervention. In addition, data collected as part of a progress evaluation can also contribute to, or form the basis for, a summative evaluation. Whereas progress evaluation is useful throughout the life of the project, it is most vital during the early stages when activities are piloted and their individual effectiveness or articulation with other project components is unknown. (NSF, 2002b.)

Summative evaluation, which collects information about outcomes and related processes, strategies, and activities that have led to them, assesses a mature project's success in reaching its stated goals. Although summative evaluation (sometimes referred to as impact or outcome evaluation) frequently addresses many of the same questions as a

progress evaluation, it does so after the project has been established and the timeframe posited for change has occurred. Summative evaluation is required by decisionmakers who may determine whether to disseminate the intervention at other sites or agencies, continue funding, increase funding, modify and try again, or discontinue. (NSF, 2002b.)

Whether an evaluation is formative or summative, typically it has six phases:

1. Develop a conceptual model of the program and identify key evaluation points.
2. Develop evaluation questions and define measurable outcomes.
3. Develop an evaluation design.
4. Collect data.
5. Analyze data.
6. Provide information to interested audiences.

(NSF, 2002b.)

Tests are considered to be one data collection technique used in project evaluations. In particular, their use is appropriate when one wants to collect data on the status of knowledge or the change in status of knowledge over time. While NSF's *2002 User-Friendly Handbook for Project Evaluation* addresses issues that are important in *choosing* a test—e.g., the extent to which the test measures knowledge, skills, or behaviors relevant to one's program—it does not address the *development* of tests for use in a project evaluation. (NSF, 2002b).

Table 1 provides the teacher and student outcomes, with subsequent data source(s), from TASC's evaluation plan:

Table 1. TASC teacher and student outcomes

	Outcome	Data Source(s)
Participating teachers will demonstrate:	1. an increase in their science content knowledge	Pre-post science content knowledge test for each curriculum unit
	2. an increase in their confidence to teach science	Surveys
	3. use of inquiry-based curriculum materials	• Surveys • Interviews • Observations
	4. improved inquiry-based teaching skills	• Surveys • Interviews • Observations
	5. an improved attitude toward science	Surveys
High school science teachers will demonstrate:	1. a need to alter their science curriculum objectives as a result of teaching students with increased knowledge and skills in science	Survey
Students of participating teachers will demonstrate:	1. knowledge of science content and skill with science process	• Curriculum unit tests by teacher • Pilot state administered tests by teacher • Portfolios of student

	Outcome	Data Source(s)
		work • Student science notebooks by teacher
	2. improvement on end-of-course test scores in science, mathematics and language arts and close academic performance gaps between student subpopulations	End-of-course exams by teacher
High school students who have been previously taught by participating teachers will demonstrate:	3. an increase in their choice preference for challenging science courses, dramatically reducing differences in course choices between student subpopulations	• High school course enrollment data • Student survey of course preference selection

(TASC, 2002b).

Table 1 documents that one of TASC's expected outcomes was participating teachers', and their students', demonstrated improvement in science content knowledge and process and that the data source of such demonstrated improvement was pre-post science content knowledge tests. Because TASC had no such tests as its data source, it contracted with the Center for Educational Research and Evaluation at the University of North Carolina—Greensboro for the development of these science tests. TASC project personnel expected to use these tests that would measure science content knowledge of participating teachers and their students to evaluate the effectiveness of the TASC teacher professional development training on the participating teachers and their students.

Additionally, TASC project personnel expected to use the test results in its annual formative evaluations to NSF to provide evidence that the project was meeting its stated program goals and project objectives according to the proposed timeline.

The final part of this chapter presents the test development process, including the standards associated with testing and measurement. The purpose of this discussion is to present an overview of the steps a test developer follows to create a test and the subsequent standards that are pertinent to the use of a test.

Test Development Process

Test development is the systematic process of producing a measure of some aspect of an individual's knowledge, skill, ability, attitudes, or other characteristics by developing items and combining them to form a test. This systematic process resulting in a test is guided by the stated purpose(s) of the test and the intended inferences to be made from the test scores. (AERA, et al., 1999).

Typically, the process of test development includes the following steps:

1. Identify the primary purpose(s) for which the test scores will be used.
2. Identify behaviors that represent the construct or define the domain.
3. Prepare a set of test specifications, delineating the proportion of items that should focus on each type of behavior identified in step 2.
4. Construct an initial pool of items. Allen and Yen (1979) suggested writing one-and-a-half to three times as many items as the final version of the test will contain in order to allow for the elimination of poor performing items.
5. Have items reviewed (and revise as needed).

6. Field-test the items on a large sample representative of the examinee population for whom the test is intended.
7. Conduct an item analysis to determine statistical properties of item scores and, when appropriate, eliminate items that do not meet pre-established criteria.
8. Design and conduct reliability and validity studies for the final form of the test.
9. Develop guidelines for administration, scoring, and interpretation of the test scores.

(Crocker & Algina, 1986).

Standards Associated with Testing and Measurement

This section presents the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing* that are associated with testing and measurement.

The Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education defines a test as "an evaluative device or procedure in which a sample of an examinee's behavior in a specified domain is obtained and subsequently evaluated and scored using a standardized process." (AERA, APA, NCME, 1999, p. 3). The *Standards for Educational and Psychological Testing*, developed by the Joint Committee, promotes the sound and ethical use of tests and provides criteria for the evaluation of tests, testing practices, and the effects of test use. (AERA, et al., 1999). Thus, it is the use of tests that is the main focus of these *Standards*.

The *Standards* are organized into three parts:

- Part I: Test Construction, Evaluation, and Documentation that includes standards for validity; reliability and errors of measurement; test development and revision; scaling, norming, and score comparability; test administration, scoring, and reporting; and supporting documentation for tests.
- Part II: Fairness in Testing that includes standards on fairness and bias; the rights and responsibilities of test takers; testing individuals of diverse linguistic backgrounds; and testing individuals with disabilities.
- Part III: Testing Applications that includes standards involving general responsibilities of test users; psychological testing and assessment; educational testing and assessment; testing in employment and credentialing; and testing in program evaluation and public policy.

(AERA, APA, NCME, 1999).

Table 2 presents the *Standards* applicable to this project. Appendix B provides the Standard and its complete statement. One common theme running through these *Standards* is documentation. A test developer is able to conform to applicable *Standards* through precise, thorough, accurate, and current documentation.

Table 2. Applicable testing standards

Topic	***Standards***
Test development	3.2; 3.3; 3.4; 3.6; 3.7; 3.8; 3.9; 3.11; 3.19; 3.20; 3.22
Validity	1.1; 1.2; 1.5; 1.6; 1.7

Topic	*Standards*
Reliability and errors of measurement	2.1; 2.2; 2.3; 2.4
Scales, norms, and score comparability	4.1; 4.2; 4.4; 4.9
Test administration	5.1; 5.5; 5.10; 5.13; 5.15; 5.16
Documentation	6.2; 6.4; 6.5; 6.7; 6.13; 6.14

Test Development

The process of developing educational tests begins with a delineation of the purpose(s) of the test and the scope of the construct or the extent of the domain to be measured (Standards 3.2, 3.11). The next phase of the process includes the development and evaluation of the test specifications (Standards 3.3, 3.6). Test specifications document the format of items, tasks, or questions; the response format or conditions for responding; the type of scoring procedures; time restrictions (if applicable); characteristics of the intended population of test takers; and procedures for administration. The third phase of test development documents the development, field testing, evaluation, and selection of the items and scoring guides and procedures (Standards 3.4, 3.7, 3.8, 3.9, 3.19, 3.22). The final phase of test development includes the assembly and evaluation of the test for operational use (Standards 3.19, 3.20). (AERA, et al., 1999).

Validity

Validity, which refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests, is the most fundamental

consideration in developing and evaluating tests. The process of validation involves gathering evidence that provides a sound scientific basis for the proposed score interpretation. Because it is the interpretations of test scores required by proposed users that are evaluated, each intended interpretation must be validated. Rather than referring to types of validity, the *Standards* refer to types of validity evidence that include evidence based on test content, on response processes, on internal structure, on relations to other variables, and on consequences of testing (Standards 1.1, 1.2, 1.5, 1.6). (AERA, et al., 1999).

Reliability and Errors of Measurement

Broadly defined, a test is a set of tasks designed to deduce, or a scale to describe, examinee behavior in a particular domain, or a system for collecting samples of an individual's work in a specified area. Coupled with the test is a scoring procedure that enables the examiner to quantify, evaluate, and interpret the behavior or work samples. Reliability refers to the consistency of such measurements when the testing procedure is repeated on a population of individuals or groups. (AERA, et al., 1999).

Even under strictly controlled conditions, an individual's responses to a set of test questions vary in their quality or character from one occasion to another. Because of this variation, and in some cases because of the subjectivity of the scoring process, an individual's obtained score and the average score of a group will always reflect at least a small amount of measurement error. In classical test theory (CTT), this error-free value is referred to as the person's true score for the test; under item response theory (IRT), a closely related concept is called an examinee's latent (unobserved) ability or trait

parameter. Thus, the hypothetical difference between an examinee's observed score on any particular measurement and the examinee's true score for the procedure is called measurement error. (AERA, et al., 1999).

Generally, errors of measurement are viewed as random and unpredictable. They limit the extent to which test results can be generalized and reduce the confidence that can be placed in any single measurement. Because measurement errors are random and unpredictable, they cannot be removed from observed scores; they can, however, be summarized. Critical information on reliability that must be reported include the identification of the major sources of error; summary statistics influencing the size of such errors; and the degree of generalizability of scores across alternate forms, scorers, administrations, or other relevant dimensions (Standards 2.1, 2.2 2.3, 2.4). (AERA, et al., 1999).

Types of Reliability

The reliability coefficient—the quantity that would be obtained if we were certain of having perfectly parallel tests—is purely a theoretical concept. Perfectly parallel tests, as defined in classical test theory, are ones where each examinee has the same true score on both forms of the test, and the error variances for the two forms are equal. In order to estimate a test's reliability, we would need the true score variance of the test; this implies we would need each individual's true score and that we do not have. Therefore, we obtain *estimates* of reliability. (Crocker & Algina, 1986).

One way we may estimate reliability is to approximate obtaining parallel measurements by administering the same form of a test on two separate occasions to the

same group of examinees. This provides an index of *test-retest* reliability between test scores and is called the *coefficient of stability*. It is an estimate of the extent to which scores are stable over a given period of time. The theory assumes that examinees' true scores do not change over the test-retest interval and that errors of measurement are entirely random. Thus, if we have two tests, test 1 and test 2, an examinee's observed score (X_1) is comprised of his/her true score (T_1) and measurement error (E_1):

$$X_1 = T_1 + E_1$$

$$X_2 = T_2 + E_2$$

The test-retest reliability coefficient reduces to: $\rho_{12} = \sigma^2_T / \sigma^2_X$, or true score variance divided by observed score variance (Crocker and Algina, 1986; Allen and Yen, 1976.)

Another way to estimate reliability is to approximate parallel measurements by administering two different forms of a test, based on the same table of specifications, on one occasion to the same examinees, and correlate the two sets of scores. This provides an index of *parallel-forms* reliability and results in a *coefficient of equivalence*. Or we could administer two alternate test forms on separate testing occasions. This provides an index of *alternate-forms* reliability, yielding a *coefficient of stability and equivalence*. As above, the derived correlation coefficient between tests reduces to σ^2_T / σ^2_X (Crocker and Algina, 1986; Allen and Yen, 1979).

Lastly, a third method for estimating reliability results in an *index of internal consistency*. This approach—internal consistency reliability—is an estimate of the extent to which all of the items measure the same construct (unobserved, or "constructed",

variable) or constructs. The index of internal consistency is a measure of the extent to which the test, taken as a whole, measures a single construct or set of highly related constructs. One practical advantage to estimating test reliability using this method as opposed to the previous two methods is that only a single administration of the test is required. (Crocker & Algina, 1986).

Methods most commonly employed for estimating the internal consistency reliability of a test include: (1) split-half reliability that divides the test into ideally parallel halves—e.g., first half/second half or odd items/even items—and correlates the scores for the two halves (adjusting for the reduced test length); and (2) Cronbach's α, that effectively divides the test into as many "mini-tests" as there are items, i.e., it is the mean of all possible split-half coefficients. (Crocker & Algina, 1986; Allen and Yen, 1979.)

Scales, Norms, and Score Comparability

Test scores are reported on scales designed to assist score interpretation, for instance, coding test questions using 0 or 1 to represent an incorrect/correct response. A raw score is obtained when all the item scores are combined. Test features—such as test length, choice of time limit, item difficulties, and the circumstances under which the test is administered—influence raw scores making them difficult to interpret in the absence of additional information. One way to facilitate interpretation and statistical analyses of raw scores is to convert them into a different set of values called derived scores or scale scores. Another way to facilitate score interpretation is to establish standards or cut

scores that distinguish different score ranges. Cut scores may be established for either raw or scale scores. (AERA, et al., 1999).

Test developers must document the construction of scales used for reporting scores (Standard 4.2); provide clear explanations of the meaning and intended interpretation of derived score scales, along with their limitations (Standard 4.1); and provide clear explanations of the meaning and intended interpretation of raw scores, along with their limitations (Standard 4.4). In addition, when scales are designed for criterion-referenced interpretation, the test developer must clearly document the rationale for the recommended score interpretations (Standard 4.9). (AERA, et al., 1999).

Test Administration, Scoring, and Reporting

Standardization of a test occurs when directions to examinees, testing conditions, and scoring procedures follow the same detailed procedures. The goal of standardization—to provide accurate and comparable measurement for everyone and unfair advantage to no one—dictates the degree of standardization along with the intended use of the test. (AERA, et al., 1999).

Test developers must document the standardized test administration and scoring procedures to be used (Standards 5.1, 5.5); the reporting of test scores, along with appropriate interpretations (Standard 5.10); the protection of the confidentiality of individually identified test scores (Standard 5.13); and data retention protocols to be followed (Standards 5.15, 5.16). (AERA, et al., 1999).

Supporting Documentation

A test's documentation is the primary means by which test developers, publishers, and distributors communicate with test users. Test documentation should be complete, accurate, current, and clearly written so that an intended reader can fully comprehend the content (Standards 6.2, 6.13, 6.14). Typically, a test's documentation specifies the nature of the test and its intended use; the processes involved in the test's development; technical information related to development; technical information related to scoring, interpretation, and evidence of validity and reliability; scaling and norming, if appropriate, to the instrument; and guidelines for test administration and interpretation. (AERA, et al., 1999).

Test documentation should include a description of the intended test-taking population along with the test specifications (Standard 6.4), statistical descriptions and analyses that provide evidence of the reliability of scores and the validity of their recommended interpretations (Standard 6.5), identification and description of instructional materials, if applicable (Standard 6.6), and qualifications of test users (Standard 6.7). (AERA, et al., 1999).

Summary

To summarize, the development of the science tests took place within a teacher-scientist collaborative project's evaluation, which project (i.e., Teachers and Scientists Collaborating project) took place within a national math-science program, which program (i.e., NSF's Math-Science Partnership program) is one of many science programs within science education reform. Applicable testing standards from the AERA/APA/NCME

1999 *Standards for Educational and Psychological Testing*—that promote the sound and ethical use of tests and provide criteria for the evaluation of tests, testing practices, and the effects of test use—were presented.

As stated previously, TASC requested the development of third, fifth, and eighth grade science tests to measure improvements in content knowledge of the TASC-trained teachers and their students. TASC project personnel intended to use the tests' results to: (1) evaluate the impact the teacher professional development training strategies and activities had on the participants (i.e., TASC-trained teachers and their students), and (2) inform NSF that the project was meeting its stated program goals and project objectives according to the proposed timeline.

The *Standards for Educational and Psychological Testing* (AERA, et al., 1999) clearly document that validity is the most fundamental consideration in developing and evaluating tests. That is, does the test measure what it purports to measure? Although there are three major types of validity—i.e., content validity, criterion-related validity, and construct validity, the type most pertinent to the immediate discussion is that of content validity.

Content validity is established through a substantive logical analysis of the content of a test with its determination based on individual, subjective judgment The two main types of content validity are face validity and logical, or sampling, validity. Face validity is established when a person examines a test and concludes that it measures the relevant trait. Logical validity involves the careful definition of the domain of behaviors

to be measured by a test and the logical design of items to cover all the important areas of this domain. (Allen & Yen, 1979).

To provide evidence of content validity, TASC had decided that item writers would be teachers who had used the NSF-approved curriculum kit on which they had received training. In addition, because one of the outcomes of NC education is that students learn to think critically and creatively, and since the NC Standard Course of Study (NCSCS) provides the competencies that students should demonstrate, all test questions were classified by two dimensions:

(1) by the NCSCS Instructional Objective (NC DPI, 2004) assessed by the question, and
(2) by the NC Thinking Skill(s)—a blend of Bloom's Taxonomy (Bloom, 1956) and Marzano's Dimensions of Learning (Marzano, et al., 1988; Marzano, 1992)—utilized by an examinee to correctly answer the question.

However, at the start of the test development process, problems developed, including:

- lack of a clear understanding as to how project effectiveness was to be measured;
- lack of communication between project personnel;
- project personnel's lack of assessment literacy;
- project personnel's lack of understanding of the test construction process; and
- project participants' lack of content knowledge and of item-writing.

Thus, the remainder of this dissertation addresses these difficulties and their impact on the actual development of the tests and suggests ways to make planned test

development and actual test development more congruent—in other words, suggestions for narrowing the gap between theory and practice in the test construction process within a project evaluation. The following chapters present the methodology (Chapter Three), the results (Chapter Four), and the conclusions and recommendations (Chapter Five) from examining the incongruence between the planned test development process and the actual test development process.

CHAPTER III

METHODOLOGY

Objective

This dissertation focuses on the test development process within a project evaluation and, specifically, on the gap between theory and practice in this process. In chapter one, the overall research question addressed by this dissertation was presented; that is, how did the actual (i.e., observed) test development process differ from the planned (i.e., expected) test development process? In Chapter Two, the context in which this study took place was presented diagrammatically as four embedded rings in Figure 2 (reproduced here). Each "ring" of potential influence was presented and discussed,

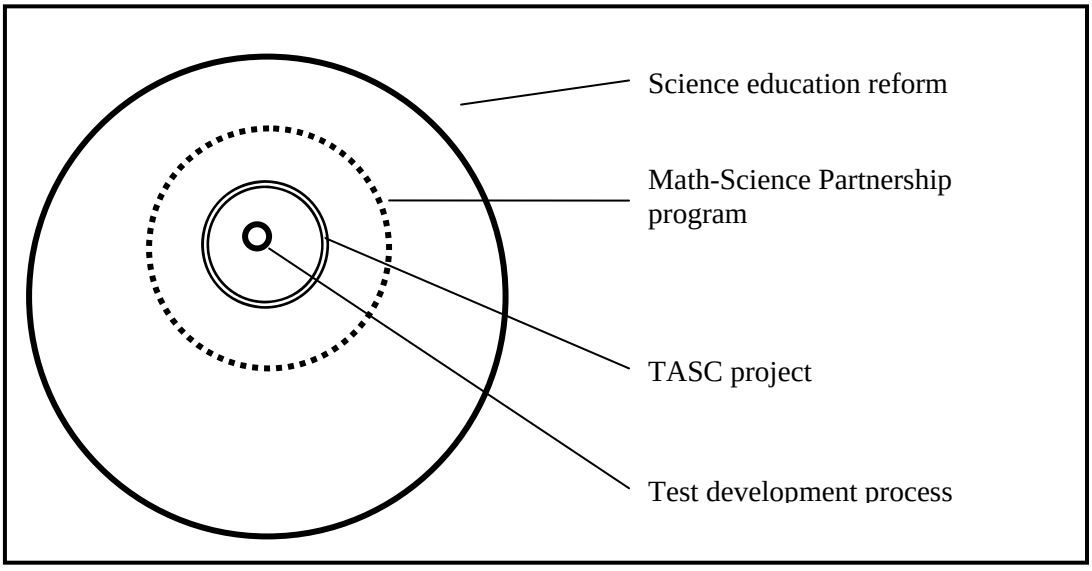


Figure 2 (from Chapter Two). Context of test development process

beginning with science education reform as reflected by the *National Science Education Standards*, then moving to one science education reform program—the National Science Foundation's Math-Science Partnership program, then moving to one MSP project—Teachers and Scientists Collaborating, and finally to the test development process itself that took place within TASC's project evaluation.

This chapter provides a detailed description of the methodology—i.e., case study—and research procedures used in the present study. The objectives of this chapter are to define the case and to present the sources of evidence and methods of analysis used to examine the gap between theoretical and actual test development process within a project evaluation.

The first section, Research Methodology, defines case study in general, presenting common characteristics of case studies. The discussion then moves to defining the "case" for this study, including a presentation of the issues (or sub-research questions) emanating from the primary research question examined by this study and explaining why case study was selected.

The second section, The Case, presents the details of the unit of analysis for this study; that is, the test development task that takes place within the MSP project evaluation. The discussion is organized around the four phases of the test construction process as presented in the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). This section addresses where the test development task fits in as part of TASC's project evaluation.

The third section, Data Sources and Acquisition, presents the sources of evidence used in this case study to document the gaps between theory and practice in the test development process. The final section, Methods of Data Analysis, presents the methods used to analyze the sources of evidence presented in the prior section.

Research Methodology

Studies related to test development frequently focus on the product of test development—the test—and use quantitative methodology. In this study, tests were developed; however, they are not the focus of the research. Rather, the focus of this study is the *process* of test development; that is, to explore and explain those factors that affected the test development process within a project evaluation. Because the natural setting in which this process occurred was the direct source of data and because the researcher—the key instrument in this research—was a participant in this process, case study was selected as the appropriate methodology (Fraenkel & Wallen, 2003).

Flyvbjerg (2006) defined a case study as an in-depth, longitudinal examination of a single instance or event—a case—that provides a systematic way of looking at events, collecting data, analyzing information, and reporting the results. From such an examination, the investigator may gain a sharpened understanding of why the instance happened as it did and what might become important to look at more extensively in future research.

Yin (2003) indicated that, generally, case study is "the preferred strategy when 'how' or 'why' questions are being posed, when the investigator has little control over events, and when the focus is on a contemporary phenomenon within some real-life

context" (p. 1). Miles and Huberman (1994) and Merriam (1998) concluded that the context in which the phenomenon occurs is a "bounded system" (borrowed from Smith (1978)) and that the case is the unit of analysis. Stake (1995) added that "the case is an integrated system" (p. 2) thereby enabling us to see the case as a thing, a single entity, a unit around which there are boundaries.

Punch (2005) provided four characteristics of case studies. First, the case has boundaries, even though the boundaries between case and context may not be clearly evident. Second, the case is a case of something, i.e., some phenomenon of interest. Third, there is an overt attempt to preserve the wholeness, unity, and integrity of the case. Fourth, multiple sources of data and multiple data collection methods are likely to be employed, typically in a naturalistic setting.

In the current research project, the case (i.e., the unit of analysis or the bounded system to be investigated) is the *test development task*, which included both process (that of developing the tests) and product (the outcome of this task—the tests). This test development task—the contemporary phenomenon of interest—was bounded by the context in which it took place and over which the researcher had no control. That is, the case took place within an evaluation of a Math Science Partnership project that, in turn, took place within the National Science Foundation's MSP program that, in turn, took place within the larger science education reform context. The case was bounded not only by the context in which it took place but also by time. For instance, this project's test development task was initiated in March 2005 and continued through September 2006—the third and four contract years of a five-year NSF MSP grant.

The purpose for studying the case—the unit of analysis—is to gain a more in-depth, holistic understanding of the real-life test development task taking place within a project evaluation context. The overriding question addressed by an investigation of this case is how the actual test development process differed from the planned test development process; or to restate it another way, what factors affected the actual test development process and how did these factors affect it? This question foreshadows differences, or problems, that emerged.

Stake (1995) referred to such problems as issues, which he described as follows:

> Issues are not simple and clean, but intricately wired to political, social, historical, and especially personal contexts. . . . Issues draw us toward observing, teasing out the problems of the case, the conflictual outpourings, the complex backgrounds of human concern. Issues help us . . . recognize the pervasive problems in human interaction. Issue questions or issue statements provide a powerful conceptual structure for organizing the study of a case. (p. 17)

Yin (2003) referred to such problems as propositions, directing attention to that which should be examined within the scope of the study.

For the current study, issues emanate from the question, How did the planned (i.e., expected, recommended) test development process differ from the actual (i.e., observed) test development process. That is, how was the actual test development process, including the resulting products, affected by:

1. national and state (i.e., NC) science standards?
2. NSF's definition of "evidence" in a project evaluation?
3. the MSP project's understanding of the role of the to-be-developed tests in their project evaluation?

4. the MSP project's understanding of the test development process? In particular, how did the MSP project's understanding of this process affect the time and money allotted for the development of these tests and their expectations?
5. the MSP project's participants (e.g., teacher item-writers and scientists)?

Case study was selected as the most appropriate means to address the issues that emerged throughout the process of developing these science tests. To have examined only the products of this process—that is, the science tests themselves—would have provided only psychometric information. Case study permits one to examine both product and process, thereby supplementing the paucity of information found in the research literature on actual test development.

The Case

The stated purpose of a test is foundational to the test development process. Tests are designed and developed for a variety of purposes. For instance, tests can be used for selection, such as selecting new employees from a group of job applications. They can be used for classification, such as classifying a student as weak in phonemic skills. Tests can also be used for evaluation, such as evaluating the effectiveness of teaching programs or professional development programs. (Allen and Yen, 1979). It is the third use of tests that is most pertinent to this study.

As stated previously in Chapter Two, formative and summative evaluations are two types of evaluation used to appraise the effectiveness of one's instructional efforts. The purpose of a formative evaluation is improvement in one's instructional program.

Summative evaluation, on the other hand, focuses on providing evidence to support decision-making. (Popham, 1999).

A test's purpose should be reflected in how test scores will be interpreted and used. Two distinct assessment strategies—norm-referenced and criterion-referenced—are used for test score interpretation. With a norm-referenced test, an examinee's test score is interpreted relative to the test scores of a larger pool of examinees. In theory, this larger pool of examinees should be similar in characteristics to the smaller group of test takers. In contrast, results from a criterion-referenced test are used to identify examinees who have, or have not, achieved certain competencies. That is, a criterion-referenced test interpretation is an absolute interpretation because it hinges on the extent to which the criterion assessment domain represented by the test is actually possessed by the examinee. Once an assessment domain is defined, an examinee's test performance can be interpreted according to the degree to which the domain has been mastered. (Popham, 1999; Hopkins, 1998.)

In this study, planned test development has been defined as the process of creating tests according to the well-established test development procedures recommended by the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing.* Here, there were really two "parts" of the planned test development process. The first "part" was the “plan” for developing these science tests for TASC—a "plan" that was articulated in the subcontract agreement between CERE/UNCG and TASC/Duke University. While CERE/UNCG had some input in this "plan", TASC/Duke University was the primary party responsible for the subaward's Scope of Work under which CERE worked. The

second "part" of the planned test development process was the professional testing standards that guided this researcher as test developer.

Because educational decision-making is at the center of educational testing, the more psychometrically sound an assessment, the more confident one can be in the decisions based on that assessment's results. In fact, the primary focus of the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing* is to promote the sound and ethical use of tests by providing criteria for the evaluation of tests, testing practices, and the effects of test use.

From Chapter Two, Table 1 presented the teacher and student outcomes from TASC's evaluation plan. Table 3, below, focuses on one of TASC's project evaluation activities, i.e., teacher and student science content and process skills. The TASC project director decided that criterion-referenced science tests, rather than norm-referenced tests,

Table 3. TASC project outcome 1

	Outcome	**Data Source(s)**
Participating teachers will demonstrate:	4. an increase in their science content knowledge	Pre-post science content knowledge test for each curriculum unit
Students of participating teachers will demonstrate:	4. knowledge of science content and skill with science process	• Curriculum unit tests by teacher • Pilot state administered tests by teacher • Portfolios of student work, • Student science notebooks by teacher

would be more appropriate as the data source with which to measure this project outcome. Because TASC did not possess such tests, in January 2005, TASC approached the Center for Educational Research and Evaluation at the University of North Carolina at Greensboro (CERE/UNCG) for the purpose of developing these tests.

The subcontract between Duke University (i.e., TASC) and CERE/UNCG was to begin in February 2005. It included the construction of pre- and post-tests to be administered to TASC-participating teachers and their students. The tests were expected to be used with NSF-approved science units, selected by TASC personnel based on their alignment with specific North Carolina science competency goals from the 2004 NC Standard Course of Study, K-8. Table 4, below, presents the ten curriculum units initially given priority by TASC project personnel, along with the matching NC competency goal(s), for which tests were expected to be created under the subcontract.

Table 4. Curriculum units and matching NC science competency goals

Grade	Curriculum Unit	NC Science Competency Goal(s)
3	Plant Growth & Development	Goal 1: The learner will conduct investigations and build an understanding of plant growth and adaptations.
3	Soils	Goal 2: The learner will conduct investigations to build an understanding of soil properties.
3	Investigating Objects in the Sky	Goal 3: The learner will make observations and use appropriate technology to build an understanding of the earth/moon/sun system.
3	Human Body	Goal 4: The learner will conduct investigations and use appropriate technology to build an understanding of the form and function of the skeletal and muscle systems of the human body.
5	Ecosystems	Goal 1: The learner will conduct investigations to build an understanding of the interdependence of plants and animals.

Grade	Curriculum Unit	NC Science Competency Goal(s)
5	Landforms	Goal 2: The learner will make observations and conduct investigations to build an understanding of landforms.
5	Investigating Weather Systems	Goal 3: The learner will conduct investigations and use appropriate technology to build an understanding of weather and climate.
5	Motion & Design	Goal 4: The learner will conduct investigations and use appropriate technologies to build an understanding of forces and motion in technological designs.
8	Earth History	Goal 5: The learner will conduct investigations and utilize appropriate technologies and information systems to build an understanding of evidence of evolution in organisms and landforms
8	MicroLife (this unit replaced the "Solutions & Pollution" unit)	Goal 6: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of cell theory. Goal 7: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of microbiology.

Chapter Two presented the *Standards for Educational and Psychological Testing* that promote the sound and ethical use of tests and provide criteria for the evaluation of tests, testing practices, and the effects of test use. (AERA, et al., 1999). Table 5, below, summarizes the four phases of the test development process referred to in the *Standards*.

Table 5. Test development process and applicable *Standards*

Standards for Educational and Psychological Testing
(AERA, APA, NCME, 1999)

Test Development Process	Standards
Phase 1: Establish the test's framework, that is: • purpose of the test and • scope of the construct (i.e., what it is to measure).	• 3.2 • 3.11
Phase 2: Develop and evaluate the test specifications, that is: • the format of items, tasks, or questions; • the response format or conditions for responding; • the type of scoring procedures; • time restrictions, if applicable; • number of items; • test blueprint: o instructional objectives to be measured and o cognitive skills to be required of examinees; • characteristics of intended test-takers; • procedures for administration	• 3.3 • 3.6
Phase 3: Construct and evaluate the initial (or pilot) test, that is: • generate items • select items based on: o content quality and scope o instructional objective addressed o cognitive skill to be used by examinee o appropriateness of the item for population of intended testtakers • assemble items into pilot test • administer pilot test to subset of intended population of test-takers • evaluate items from piloting test (i.e., item analysis) • evaluate scoring procedures • evaluate test administration procedures	• 3.4 • 3.7 • 3.8 • 3.9 • 3.19 • 3.22

Standards for Educational and Psychological Testing
(AERA, APA, NCME, 1999)

Test Development Process	Standards
Phase 4: Assemble and evaluate test for operational use: o revise, replace, or delete items based on pilot test results o assemble items for operational test o revise test administration procedures, if applicable o revise scoring procedures, if applicable o administer operational test to intended population of test-takers o evaluate operational test results	• 3.19 • 3.20

Table 6 presents the timeline incorporated in the TASC-CERE subcontract for the accomplishment of the test development task. This table documents TASC's expectation that this task would be completed within about ten months.

Table 6. TASC-CERE subcontract deliverables

DATE	DELIVERABLE
Feb 2005	Item writing workshops
Mar – Dec 2005*	Write & develop 32 modules of test items (based on science kits carried by TASC)
Mar – Apr 2005	Pilot 1st 5 modules and analyze items (students & teachers)
Apr – May 2005	Pilot 2nd 5 modules and analyze items, revise 1st 5 as needed (student & teachers)
May – Jun 2005	Pilot 3rd 5 modules and analyze items, revise 2nd 5 as needed (student & teachers)
mid-Jun 2005	A report, to be included in TASC's annual report to NSF, on preliminary analysis of student and teacher changes in knowledge and

DATE	DELIVERABLE
	science process skills, pre- and post-.
Jun – Jul 2005	Pilot 4th 5 modules and analyze items, revise 3rd 5 as needed (student & teachers)
	Pilot 7 more modules and analyze items, revise 4th 5 as needed (student & teachers)
Sep 2005	Complete analysis and revision of all modules
Oct 2005*	Statistical analysis and Q Matrix to identify skill mastery

Phase 1: Test Framework

The purpose of each test was to enable TASC to evaluate the instructional effectiveness of its teacher professional development through demonstrated improvement in teachers', and their students', science content and process knowledge. TASC planned to use results from each test to ascertain (1) whether teachers, who had received TASC training on a particular curriculum unit, demonstrated improvement in their science content and process knowledge, and (2) whether these teachers' students demonstrated improvement in science content and process knowledge from their teachers' training and use of the curriculum unit.

Phase 2: Test Specifications

A table of specifications—or blueprint—is the foundation for establishing content-related evidence of validity for a test. The blueprint guides the selection of test questions that reflect achievement of the content and course objectives. The blueprint answers the question: What is being measured? (McDonald, 1999).

Once a set of objectives has been chosen, there needs to be a plan for deciding the relative emphasis each objective should receive on the test. Specifically, there should be a balance of items so that different components of the construct are represented in proportion to their perceived importance. (Crocker and Algina, 1986).

In the context of this study, the table of specifications was a two-dimensional grid where each row represented competencies in the particular content area and each column represented the cognitive processes utilized. The competencies represented were derived from the science competency goals with the accompanying instructional objectives from the 2004 North Carolina Standard Course of Study. The taxonomy of cognitive processes used was the NC Thinking Skills, the taxonomy adopted by the NC Department of Public Instruction, one of the initial partners in the TASC project, to classify questions for North Carolina tests. Table 7, which presents common taxonomies used to classify instructional objectives, demonstrates that the NC Thinking Skills model appears to be a blend of Bloom's Taxonomy (Bloom, 1956) and Marzano's Dimensions of Learning (Marzano, 1988).

Table 7 Taxonomies used to classify instructional objectives

Bloom's Taxonomy	**Marzano's Dimensions of Learning**	**NC Thinking Skills**	**Description of NC Thinking Skills** (from *North Carolina Thinking Skills: An introduction* by Tom Munk (Oct. 2001))
knowledge	focusing information-gathering remembering	knowledge	"At the lowest level, students should learn the focusing, information-gathering, and remembering skills that allow them to gain declarative and procedural **knowledge**."

Bloom's Taxonomy	**Marzano's Dimensions of Learning**	**NC Thinking Skills**	**Description of NC Thinking Skills** (from *North Carolina Thinking Skills: An introduction* by Tom Munk (Oct. 2001))
comprehension	organizing	organizing	"Techniques such as comparing, classifying, ordering, and representing allow students to develop skill in **organizing** information."
application		applying	"**Applying** their knowledge to a novel situation is a higher-level skill that our children will need to succeed, both in school and outside the classroom."
analysis	analyzing	analyzing	"By examining the parts and relationships of existing information, students clarify their knowledge and practice the learning skill of **analyzing**."
	generating	generating	"By inferring, predicting, and elaborating, students can become skilled at **generating** new information, meaning, or ideas."
synthesis	integrating	integrating	"**Integrating** can be accomplished by condensing information efficiently into a cohesive statement or by connecting existing and prior knowledge into a new understanding."
evaluation	evaluating	evaluating	"**Evaluating** ideas by setting criteria and confirming the accuracy of claims is the last of the North Carolina Thinking Skills."

At this researcher's request, TASC personnel estimated, for each workshop, the percentage of training time spent on each of the applicable instructional objectives. These percentages were TASC's best estimation of training times; that is, TASC personnel had not objectively validated either the actual amount of training time or the actual amount of teachers' instructional time spent on each instructional objective. Appendix C presents the test blueprints that include the NC competency goal, instructional objectives, the estimated percentage of TASC training time spent on the instructional objectives, the minimum number of items to be created for each objective, and the NC Thinking Skills cognitive processes.

The tests would be used *within* a program evaluation context—i.e., within the TASC project. That is, TASC expected to use the tests to evaluate the instructional effectiveness of the professional development provided by its trainers to those teachers in one of the four school districts that initially collaborated in the TASC project. TASC anticipated being able to "fine tune" its professional development training based on results from the tests.

Phase 3: Pilot Test

Item Development

The DUKE/TASC-UNCG/CERE subcontract stipulated that CERE would conduct an item writing workshop because TASC expected that its trained teachers would write the items for these tests. In addition, TASC determined that each test—the teacher test and the student test—would include 10 to 20 multiple-choice items, none of which would be knowledge-level questions.

Recruitment of teachers.

TASC criteria for selecting potential item-writers included: (1) teachers who had received prior TASC training on at least one science curriculum workshop, (2) TASC-trained teachers who had used the science unit in their classrooms, *and* (3) TASC-trained teachers who attended one three-hour item writing workshop, prepared and presented by this researcher, at the TASC Training Center in Durham, NC. In addition, TASC stipulated two item writers per science unit.

Item writing workshop.

The subcontract between TASC and CERE stipulated an item writing workshop be conducted by CERE. It was expected that the workshop would last approximately three hours and that all item writers would attend. The workshop, to be attended by the teacher item-writers and by the TASC project director, would present a brief overview of the TASC project and the test development process, as well as a strong emphasis on the use of the to-be-developed tests. It would present detailed information regarding the construction of multiple choice items, especially those requiring the use of higher thinking skills, and the submission of the items using an item specification sheet (Appendix D). A copy of the *Multiple Choice Item Writing Workbook*, prepared by the author and used for the item writing workshop, is included as Appendix E.

Item generation and revisions.

After attending the item writing workshop, teacher item-writers would be instructed to submit 16 multiple choice content questions—8 teacher-test items and 8 student-test items, all written above the knowledge-level category. Item-writers would be

given two weeks to submit draft questions, each question submitted with an item specification sheet, to this researcher who would review each question submitted and request revisions.

Test Construction

Final versions of items would be assembled into initial draft tests—i.e., pilot tests—that would be submitted to TASC for its review and revision, if necessary. The tests would then be prepared for administration to a subset of the targeted population of teacher and student test-takers.

Administration Procedures

Standardized test-taking instructions, including the recording of answers onto the answer sheet provided along with the test, would be incorporated on the first page of the tests. Standardized test administration instructions sheets would be provided for TASC workshop instructors who would administer the teacher pretests and for the TASC-trained teachers who would administer the pretests and posttests to their students.

Teacher pretests would be administered by TASC workshop instructors at the beginning of Session 1 of the workshops. Student pretests would be administered by TASC-trained teachers to their students prior to use of the science units. Teacher posttests would be self-administered at the end of the science units during the time when the teachers administered the posttests to their students.

All teacher and student pretests and posttests would be returned to TASC along with the science units. TASC would then have the answer sheets (i.e., Scantrons) delivered to this researcher for processing.

Scoring Procedures

Once received at CERE, answer sheets would be submitted to the University Teaching and Learning Center (TLC) at UNCG for processing. TLC would scan the answer sheets and create data files that included each examinee's unique identification number, gender, grade, and recorded responses (i.e., A, B, C, etc.) to each test question. After the data files were received from TLC, the content questions would be recoded—using statistical software (i.e., SPSS)—as correct or incorrect according to the answer keys provided by TASC personnel. Examinees' identities would be protected through the use of unique identification numbers.

Pretest and posttest data would be matched and scores—aggregated by instructional objective—would be reported to TASC.

Pilot Testing, Item Analyses, and Test Revision

As previously presented in Table 6, the TASC-CERE subcontract stipulated that pilot testing of student and teacher pretests and posttests, item analyses, and test revisions would occur from April through September 2005 and final reporting would occur by October 2005.

Phase 4: Operational Test

There were no provisions in the original TASC-CERE subcontract for any of the Phase 4 activities presented in Table 5.

Data Sources and Acquisition

Sources of evidence most commonly used in case studies include documentation, archival records, interviews, direct observations, participant-observation and physical artifacts. (Yin, 2003). One strength of such qualitative evidence is that the focus is on naturally occurring, ordinary events in natural settings. Data are collected in close proximity to a specific situation, rather than through the mail or over the phone. The emphasis is on a particular case, a focused and bounded phenomenon embedded in its context, with such influences of local context taken into account. The potential for understanding latent, underlying, or nonobvious issues is strong. (Miles & Huberman, 1994).

A second strength of qualitative data is their richness and holistic nature, with strong potential for revealing complexity. Such data provide vivid descriptions, nested in a real-life context. In addition, such data are typically collected over a sustained period of time, making them powerful for studying any process. (Miles & Huberman, 1994).

Lastly, qualitative data often have been advocated as the best strategy for discovery, exploring a new area, developing hypotheses. Such data have strong potential for testing hypotheses, i.e., seeing whether specific predictions hold up. In addition, qualitative data are useful when one needs to supplement, validate, explain, illuminate, or reinterpret quantitative data gathered from the same setting. (Miles & Huberman, 1994).

For the current study, qualitative data sources included:

- documentation, such as:

- contractual documents (NSF Project Solicitation; TASC proposal to NSF; TASC project evaluation plan, including logic model; UNCG subcontracts; TASC annual reports to NSF),
- personal communications (electronic mail) with TASC personnel,
- investigator notes; and

- interviews of TASC personnel (Project Director, Project Evaluator, Training Director, Professional Development Materials Manager, Curriculum Director, TASC-trained teacher item writer(s)).

Interviewees were TASC personnel who participated in the creation of the science tests as part of the TASC project evaluation. Interviewees participated either by helping to write, evaluate, revise, and/or score test items; to distribute pilot tests; to administer pilot tests to teachers participating in the TASC project; or by anticipating the use of these tests in evaluating the TASC project. An interview protocol was developed to examine how the actual test development process was affected by TASC personnel's understanding of the role of the to-be-developed tests in the project evaluation, their understanding of the test development process and how this affected the time and money allotted for the development of these tests as well as their expectations, their work with teacher-item writers, and their work as item-writers. Quantitative data sources included analyses of the piloted science tests.

Table 8 presents the research questions investigated in this study, the affected phase(s) of the test development process, and the data sources used to document the

various sources of influence on the test development process within the project evaluation.

Table 8. Research questions and data sources

Research Question: *How was actual (i.e., observed) test development affected by*:	**Affected Phases of Test Construction Process**	**Data Source**
1. national and state (i.e., NC) science standards?	• Phase 1 (Test Framework) • Phase 2 (Test Specifications)	• *NSES* • NC Standard Course of Study
2. NSF's definition of "evidence" in a project evaluation?	• Phase 1	• NSF MSP Program Solicitation
3. the MSP project's understanding of the role of the to-be-developed tests in their project evaluation?	• Phase 1 • Phase 2	• TASC proposal to NSF • TASC evaluation plan • DUKE/TASC subaward with UNCG/CERE • Test blueprint • Researcher interview with Project Director • Researcher interview with Project Evaluator
4. the MSP project's understanding of the test development process? In particular, how did the MSP project's understanding of this process affect the time and money allotted for the development of these tests and their expectations?	• Phase 2 • Phase 3 (Pilot Test) • Phase 4 (Operational Test)	• DUKE/TASC subaward with UNCG/CERE • Investigator notes • Researcher's personal communications (emails) with project personnel • Researcher's interviews with TASC Project Director • Researcher's interview

Research Question: *How was actual (i.e., observed) test development affected by*:	**Affected Phases of Test Construction Process**	**Data Source**
		with Project Evaluator
5. the MSP project's participants (e.g., teacher item-writers and scientists)?	• Phase 3	• Researcher's interviews with TASC project personnel • Researcher's personal communications (i.e.,emails) with teacher item-writers • Originally submitted items from teacher item-writers

Methods of Data Analysis

Miles and Huberman (1994) defined qualitative data analysis as consisting of three concurrent flows of activity: data reduction, data display, and conclusion drawing/verification. Data reduction is defined as the process of selecting, focusing, simplifying, abstracting, and transforming the data that appear in written-up field notes or transcriptions. Data reduction is not separate from analysis; rather, it is part of analysis. It is a form of analysis that sharpens, sorts, focuses, discards, and organizes data in such a way that "final" conclusions can be drawn and verified. By data reduction, the authors noted that they did not necessarily mean quantification. Rather, they stated that qualitative data can be reduced and transformed in many ways, e.g., through selection, through summary or paraphrase, through being subsumed in a larger pattern, etc. They

emphasized the importance of not stripping the data at hand from the context in which they occurred.

Miles and Huberman (1994) generically defined display as "an organized, compressed assembly of information that permits conclusion drawing and action." (p. 11). It is a way to help the researcher understand what is happening and to act based on that understanding. Designing a display includes deciding on the rows and columns of a matrix for qualitative data and deciding which data, in which form, should be entered in the cells—all of which are analytic activities.

Lastly, Miles and Huberman (1994) described conclusion drawing and verification. They noted that from the start of data collection, the qualitative analyst decides what things mean by noting regularities, patterns, explanations, possible configurations, causal flows, and propositions. These initial conclusions are held lightly, with the researcher maintaining openness and skepticism. Increasingly, conclusions become more explicit and grounded as they are verified by the researcher; that is, meanings derived from the data have been tested for their plausibility, their sturdiness, their confirmability—i.e., their validity.

Bogdan and Biklen (1992) stated that data analysis is the process of systematically searching and arranging the interview transcripts, field notes, and other materials that a qualitative researcher has accumulated to increase his/her understanding of them and to enable him/her to present what he/she has discovered to others. Analysis involves working with data, organizing them, breaking them into manageable units,

synthesizing them, searching for patterns, discovering what is important and what is to be learned, and deciding what to present to others.

Data analysis includes examining, categorizing, tabulating, testing, and otherwise recombining both quantitative and qualitative evidence to address the initial issues or propositions of a study. Given that analyzing case study evidence is particularly difficult because the strategies and techniques have not been well defined, a general analytic strategy is that of defining priorities for what to analyze and why. Three strategies are:

- relying on theoretical propositions upon which the original objectives and design of the case study were based,
- establishing a framework based on rival explanations, and
- developing case descriptions.

(Yin, 2003).

While any of these strategies can be used in practice, five specific techniques for analyzing case studies include:

- pattern matching, which compares an empirically based pattern with a predicted one (or with several alternative predictions);
- explanation building—a variation of pattern matching—where the goal is to analyze the case study data by building an explanation about the case;
- time-series analysis, which is directly analogous to the time-series analysis conducted in experiments and quasi-experiments;

- logic models, which deliberately stipulate a complex chain of events over time and match empirically observed events to theoretically predicted ones; and
- cross-case synthesis, particularly relevant where two or more cases are included in the study.

Regardless of the choice of strategies or techniques, a constant challenge is that of producing high-quality analyses, which require researchers to attend to *all* the evidence, display and present evidence apart from any interpretation, and show adequate concern for exploring alternative interpretations. (Yin, 2003).

Computer-assisted routines with prepackaged software—e.g., Nonnumerical Unstructured Data Indexing, Searching, and Theorizing (NUD*IST; Gahan & Hannibal, 1999) or Computer Assisted Qualitative Data Analysis Software (CAQDAS; Fielding & Lee, 1998)—have become increasingly popular. The software enables a researcher to code and categorize large amounts of narrative text, for instance, that which have been collected from open-ended interviews or from historic documents. (Yin, 2003).

The great benefit from such tools is when the narrative texts represent a *word for word* record of an interviewee's remarks or the literal content of a file or historic document, and the empirical study is trying to derive meaning and insight from the word usage and frequency pattern found in the texts. However, these verbatim or documentary records are typically only part of one's total case study. There remains the need for an analytic strategy to address the larger or fuller case study. (Yin, 2003).

Miles and Huberman (1994) described and summarized another set of helpful analytic manipulations, which include:

- putting information into different arrays;
- making a matrix of categories and placing the evidence within such categories;
- creating data displays—flowcharts and other graphics—for examining the data;
- tabulating the frequency of different events;
- examining the complexity of such tabulations and their relationships by calculating second-order numbers such as means and variances; and
- putting information in chronological order or using some other temporal scheme.

One technique, mentioned above, for analyzing case studies is pattern matching. Yin (2003) defined pattern matching as comparing an empirically based pattern with a predicted one. Trochim (1989) defined it as an attempt to link two patterns—a theoretical pattern, or that which is expected in the data, and an observed one, or that which was actually observed in the data. Yin (2003) further indicated that a variation of pattern matching—explanation building—can be used where the goal is to analyze the case study data by building an explanation about the case. Trochim (1989), in addressing the question of how one best develops the theoretical pattern for a particular study, stated there is no one correct form which a theoretical pattern must take. That is, a pattern may

be verbal in nature, be a collection of mathematical formulae, or consist of a pictorial representation.

Because this dissertation focuses on the gap between theory and practice, pattern matching will be used to compare an expected, or theoretical, pattern to the observed, or actual, pattern in the test development task. One such theoretical pattern is presented in Table 9. This pattern represents the expected pattern for the item generation task of phase 3 (construction and evaluation of the pilot tests) of the test construction process. This expected, or theoretical, pattern was compared to the actual pattern observed from the data.

Table 9. Theoretical pattern for item generation task

Resources	Activities	Outputs	Outcomes
• Testing standards • CERE subcontract with Duke U. • TASC Institute teachers • TASC scientists • Curriculum units • NC Standard Course of Study	• MC Item Writing Workshop → MC Item Writing Workbook	• Higher order (i.e., above knowledge level) MC Qs: o 8 usable MCQs for teacher test o 8 usable MCQs for student test	• Teacher tests of 10-15 MCQs • Student tests of 10-15 MCQs

Another expected pattern was that from the pilot test results. That is, test-takers were expected to score low on a particular construct (e.g., motion and design) prior to instruction and then to score high on that same construct after instruction. Again, this theoretical pattern was compared to the actual pattern derived from item analyses data.

Punch (2005) stated that if one wants to know why something happens, it is important to have a good description of exactly what happens; and that when we know why, or how, something happens, it puts us in a position to predict what will happen or perhaps to be able to control what will happen. In this study, comparisons of expected and observed patterns were made in order to provide evidence of what occurred and why it occurred the way it did. In other words, the goal was to analyze the case study data in order to understand not merely what happened but why it happened as it did.

CHAPTER IV

RESULTS

This chapter begins with an introduction that briefly reviews the case. It then moves to the planned test development process, defined as 1) the contractual basis for developing the tests, and 2) the steps, and standards, followed by this researcher as test developer. The discussion then moves to the actual test development process, or what actually occurred as this test developer attempted to follow "the plan" for developing these tests. This part of the discussion is organized around the four phases of test development—the test framework, the test specifications, the pilot test, and the operational test—ending with a discussion of the factors that affected the particular test development phase.

Introduction

In the current research project, the case (i.e., the unit of analysis or the bounded system to be investigated) is the *test development task* that includes both process (that of developing the tests) and product (the outcome of this task—the tests). This test development task—the contemporary phenomenon of interest—was bounded by the context in which it took place and over which the researcher had no control. That is, the case took place within an evaluation of a Math Science Partnership project that, in turn, took place within the National Science Foundation's MSP program that, in turn, took

place within the larger science education reform context. The purpose for studying the case—the unit of analysis—is to gain a more in-depth, holistic understanding of the real-life test development task taking place within a project evaluation context.

This context was presented diagrammatically in Chapter Two as four embedded rings (Figure 2, reproduced here). Each "ring" of potential influence was presented and discussed, beginning with science education reform as reflected by the *National Science Education Standards*, then moving to one science education reform program—NSF's Math-Science Partnership program, then moving to one MSP project—Teachers and Scientists Collaborating, and finally to the test development task itself that took place within TASC's project evaluation.

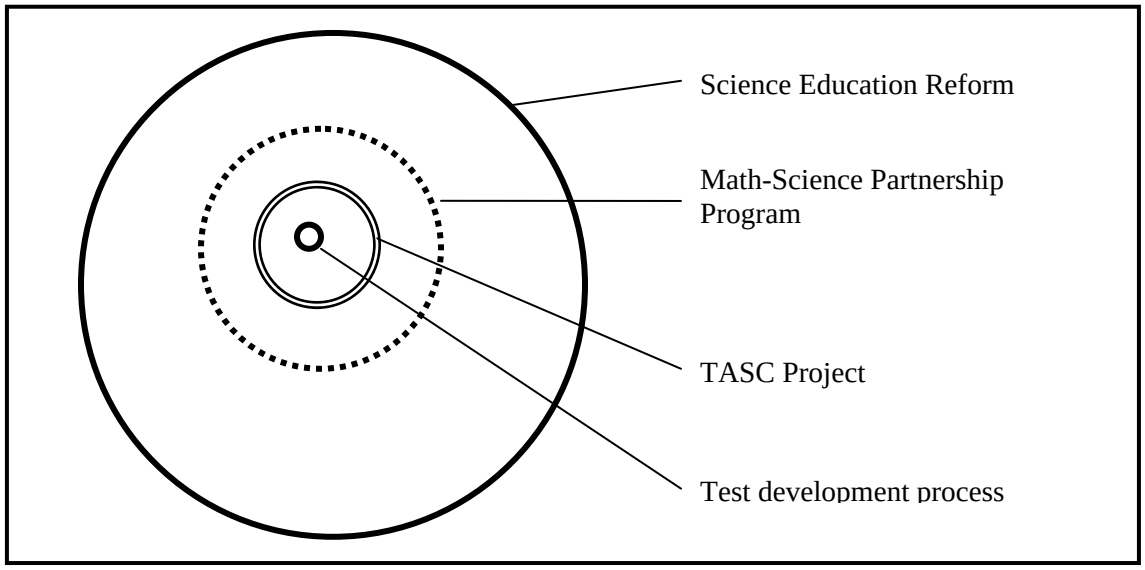


Figure 2 (from Chapter Two). Context of test development process

The overriding question addressed by the investigation of this case is how the actual test development process differed from the planned test development process; or to

restate it another way, what factors affected the actual test development process and how did these factors affect it? That is, how was the actual test development process, including the resulting products, affected by:

1. the national and state (i.e., NC) science standards?
2. the NSF's definition of "evidence" in a project evaluation?
3. the MSP project's understanding of the role of the to-be-developed tests in their project evaluation?
4. the MSP project's understanding of the test development process? In particular, how did the MSP project's understanding of this process affect the time and money allotted for the development of these tests and their expectations?
5. the MSP project's participants (e.g., teacher item-writers and scientists)?

The focus of this dissertation is documenting how various "rings of influence" affected the test development *task*, which took place within a MSP project evaluation, and how this influence on the test development task in turn affected the data collection process required by professional development providers not only to inform revisions to their programs but also to provide evidence to their funders of their program's effectiveness. To reiterate, even though TASC is the "center of attention" for this dissertation, it is merely an example of what may have become the common practice in the evaluation of science education programs.

As cited previously, data analysis is the process of systematically searching and arranging the interview transcripts, fieldnotes, and other materials that a qualitative researcher has accumulated to increase his/her understanding of them and to enable

him/her to present what he/she has discovered to others. Analysis involves working with data, organizing them, breaking them into manageable units, synthesizing them, searching for patterns, discovering what is important and what is to be learned, and deciding what to present to others. (Bogdan and Biklen,1992).

Because this study focused on the gap between theory and practice, pattern matching was widely used to compare expected, or theoretical, patterns to the observed, or actual, patterns in the test development task. After a brief review of the planned test development process from Chapter Three, this chapter then will present the actual test development process using a chronological framework within each phase. Each phase concludes with the identification of the factors that influenced it.

In this study, planned test development has been defined as the process of creating tests according to the well-established test development procedures recommended by the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing.* When developing tests for a "client", there are really two "parts" of the planned test development process. The first "part" of this process is dictated by the contractual agreement between parties. The second "part" of this process consists of the steps articulated by the professional testing standards and followed by the test developer.

For this study, the first part of the planned test development process was determined by the subcontract agreement between CERE/UNCG and TASC/Duke University. Although CERE/UNCG had some input in this plan, TASC/Duke University remained the primary party—the "client" or "customer"—responsible for the subaward's scope of work under which CERE worked. The second part of the planned test

development process consisted of the steps recommended by professional testing standards that guided this researcher as test developer.

Because educational decision-making is at the center of educational testing, the more psychometrically sound an assessment, the more confident one can be in the decisions based on that assessment's results. In fact, the primary focus of the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing* is to promote the sound and ethical use of tests by providing criteria for the evaluation of tests, testing practices, and the effects of test use.

Test Framework (Phase 1)

In the first phase of test development, the developer must establish the test's framework; that is, the purpose of the test, and the scope of the construct (i.e., what is to be measured).

Under its NSF Grant No. EHR-0227035, the Teachers and Scientists Collaborating (TASC) project at Duke University has been providing teacher trainings focused on science content, inquiry-based teaching, and effective use of science materials (i.e., NSF-approved science curriculum kits) along with support from scientists trained in helping teachers use the curriculum units. TASC's five-year grant, issued under NSF's Math-Science Partnership (MSP) program that is included in the 2001 No Child Left Behind Act legislation, began October 1, 2002 and ends September 30, 2007.

One of TASC’s anticipated impacts included a demonstrated improvement in science content knowledge of participating K-8 teachers and their students. This particular outcome, from TASC's evaluation plan, was documented by Table 3

(reproduced below) from Chapter Three. Thus, TASC expected to use test results as one source of evidence to demonstrate that the TASC teacher training on the curriculum units resulted in improved science content knowledge of participating teachers and their students.

Table 3 (from Chapter Three). TASC project outcome 1

	Outcome	Data Source(s)
Participating teachers will demonstrate:	1. an increase in their science content knowledge	Pre-post science content knowledge test for each curriculum unit
Students of participating teachers will demonstrate:	1. knowledge of science content and skill with science process	• Curriculum unit tests by teacher • Pilot state administered tests by teacher • Portfolios of student work, • Student science notebooks by teacher

To evaluate this impact, i.e., improvements in science content knowledge of participating teachers and their students, TASC/Duke University needed science content tests. In August 2004, toward the end of its second contract year, TASC contacted UNCG's Center for Educational Research and Evaluation—with many years' experience in educational research, measurement, and evaluation—by email inquiring as to its availability to work "with DPI to do, and be paid for, part of the TASC evaluation." Five months later, in mid-January 2005, TASC emailed to CERE a list of science module topics "as listed in the NC Standards" that "we'd like tested but we understand that there

are too many to test right away. We'll need to chose. To start that process, please look these over and let us know which ones would be easiest to address soonest." In TASC's email, the following modules were listed:

- Third grade:
  - Human Body (NCSCOS Goal 4, subgoals 1 through 5)
  - Plant Growth & Development (NCSCOS Goal 1, subgoals 1 through 6)
  - Soils (NCSCOS Goal 2, subgoals 1 through 6);
- Fourth grade:
  - Food Chemistry (NCSCOS Goal 4, subgoals 1 through 5)
  - Magnetism & Electricity (NCSCOS Goal 3, subgoals 1 through 9);
- Fifth grade:
  - Investigating Weather Systems (NCSCOS Goal 3, subgoals 1 through 6)
  - Landforms (NCSCOS Goal 2, subgoals 1 through 7)
  - Motion & Design (NCSCOS Goal 4, subgoals 1 through 7);
- Sixth grade:
  - Exploring Energy (NCSCOS Goal 2, subgoals 1 and 3; Goal 6, subgoals 1,2, 4, 6, 7)
  - Planetary Science (NCSCOS Goal 1, subgoals 1 through 6, 8; Goal 2, subgoals 1, 3; Goal 5, subgoals 1 through 6);
- Seventh grade:
  - Human Body Systems (NCSCOS Goal 1, subgoals 1 through 6, 8; Goal 2, subgoals 1, 3; Goal 4, subgoals 1 through 5, 8)
  - Thrill Ride (NCSCOS Goal 1, subgoals 1 through 6, 8; Goal 2, subgoals 1, 3; Goal 6, subgoals 1 through 6); and
- Eighth grade:
  - Earth History (NCSCOS Goal 1, subgoals 1 through 6, 8; Goal 2, subgoals 1, 3; Goal 5, subgoals 1 through 5)
  - Solutions & Pollution (NCSCOS Goal 1, subgoals 1 through 5, 8; Goal 3, subgoals 1, 7, 8; Goal 4, subgoals 4, 5).

CERE responded to TASC with a draft proposal for the measurement component of the evaluation for the TASC program. In CERE's draft proposal to TASC, an approximately 12-month period of performance beginning early February 2005, was projected to complete the following tasks:

- conduct item writing workshops for master teachers selected by TASC personnel to create items for 23 science modules;
- oversee the item writing (i.e., modifying and editing items);
- create student and teacher assessments;
- pilot items; and
- conduct item/test analyses.

The first "part" of the planned test development process, that is, the plan for creating these tests, was derived from the scope of work from the subaward between CERE/UNCG and TASC/Duke University. Under the scope of work, quoted from below, Duke University contracted with CERE/UNCG to:

> develop tests to measure improvements in content knowledge of participating teachers and their students.
>
> To that end, CERE will develop:
>
> - pre- and post-tests <u>for students</u> on content and science process related to kits selected from the following, giving priority to the following modules grades 3, 5, and 8:
>   - Grade 3
>   - Soils
>   - Plant Growth & Development
>   - Human Body 3
>   - Investigating Objects in the Sky
>   - Grade 5
>   - Investigating Weather Systems
>   - Motion and Design
>   - Landforms
>   - Grade 8
>   - Solutions and Pollution
>   - Earth History

        - Micro-Life
        - Ecosystems

- pre- and post-tests for teachers on content and science process related to kits selected from the following, giving priority to the following modules grades 3, 5, and 8:
    - Grade 3
        - Soils
        - Plant Growth & Development
        - Human Body 3
        - Investigating Objects in the Sky
    - Grade 5
        - Investigating Weather Systems
        - Motion and Design
        - Landforms
    - Grade 8
        - Solutions and Pollution
        - Earth History
        - Micro-Life
        - Ecosystems

- Scoring and analysis of the above tests for students in the 4 targeted districts. These are students in about 370 classrooms in grades 3, 5, and 8 (about 6,825 students grades K-5 and about 9,215 students grades 6-8).

Payments beyond 50% of the subaward amount … are contingent upon the TASC Director's approval. The 32 modules in the table below refer to the 32 curriculum units that TASC provides its partners (as listed on the TASC web site). The TASC Director is the primary liaison. CERE assumes responsibility for the following deliverables by the dates below.

DATE	DELIVERABLE
Feb 2005	Item writing workshops
Mar – Dec 2005*	Write & develop 32 modules of test items (based on science kits carried by TASC)
Mar – Apr 2005	Pilot 1st 5 modules and analyze items (students & teachers)
Apr – May 2005	Pilot 2nd 5 modules and analyze items, revise 1st 5 as needed (student & teachers)

May – Jun 2005	Pilot 3rd 5 modules and analyze items, revise 2nd 5 as needed (student & teachers)
mid-Jun 2005	A report, to be included in TASC's annual report to NSF, on preliminary analysis of student and teacher changes in knowledge and science process skills, pre- and post-.
Jun – Jul 2005	Pilot 4th 5 modules and analyze items, revise 3rd 5 as needed (student & teachers)
	Pilot 7 more modules and analyze items, revise 4th 5 as needed (student & teachers)
Sep 2005	Complete analysis and revision of all modules
Oct 2005*	Statistical analysis and Q Matrix to identify skill mastery

*Date extends beyond the period of performance for this subcontract, which is 9/30/05.
(TASC, 2005).

The second "part" of the planned test development included the steps, and application of the *Standards*, that guided this researcher as test developer in the creation of these tests for TASC. Typically, the process of test development requires one to:

1. Identify the primary purpose(s) for which the test scores will be used.
2. Identify behaviors that represent the construct or define the domain.
3. Prepare a set of test specifications, delineating the proportion of items that should focus on each type of behavior identified in step 2.
4. Construct an initial pool of items.
5. Have items reviewed (and revise as needed).
6. Field-test the items on a large sample representative of the examinee population for whom the test is intended.

7. Determine statistical properties of item scores and, when appropriate, eliminate items that do not meet pre-established criteria.
8. Design and conduct reliability and validity studies for the final form of the test.
9. Develop guidelines for administration, scoring, and interpretation of the test scores.

The *Standards* applicable in developing these science tests were included in Table 5 (reproduced below) from Chapter Three. Each phase was discussed in Chapter Three.

Table 5 (from Chapter Three). Test development process and applicable *Standards*

Standards for Educational and Psychological Testing
(AERA, APA, NCME, 1999)

Test Development Process	**Standards**
Phase 1: Establish the test's framework, that is: • purpose of the test and • scope of the construct (i.e., what it is to measure).	• 3.2 • 3.11
Phase 2: Develop and evaluate the test specifications, that is: • the format of items, tasks, or questions; • the response format or conditions for responding; • the type of scoring procedures; • time restrictions, if applicable; • number of items; • test blueprint: o instructional objectives to be measured and o cognitive skills to be required of examinees; • characteristics of intended test-takers; • procedures for administration	• 3.3 • 3.6
Phase 3: Construct and evaluate the initial (or pilot) test, that is: • generate items • select items based on:	• 3.4 • 3.7 • 3.8 • 3.9

Standards for Educational and Psychological Testing
(AERA, APA, NCME, 1999)

Test Development Process	Standards
o content quality and scope o instructional objective addressed o cognitive skill to be used by examinee o appropriateness of the item for population of intended testtakers • assemble items into pilot test • administer pilot test to subset of intended population of test-takers • evaluate items from piloting test (i.e., item analysis) • evaluate scoring procedures • evaluate test administration procedures	• 3.19 • 3.22
Phase 4: Assemble and evaluate test for operational use: o revise, replace, or delete items based on pilot test results o assemble items for operational test o revise test administration procedures, if applicable o revise scoring procedures, if applicable o administer operational test to intended population of test-takers o evaluate operational test results	• 3.19 • 3.20

Thus, CERE/UNCG was contracted by TASC/Duke to develop tests to measure improvements in content knowledge of participating teachers and their students. The purpose of each test—the first step in developing a test—was to enable TASC to evaluate the instructional effectiveness of its teacher professional development through demonstrated improvement in teachers', and their students', science content and process knowledge. That is, TASC planned to use results from each test to demonstrate to its

funding agency (NSF) whether teachers, who had received TASC training on a particular curriculum unit, demonstrated improvement in their science content and process knowledge and whether these teachers' students demonstrated improvement in science content and process knowledge from their teachers' training and use of the curriculum units.

Factors that Affected Phase 1 (Test Framework)

Factors that affected the test framework—the purpose of the test and the scope of the construct—include national and state science standards, NSF's definition of "evidence" in a project evaluation, and the TASC project personnel's understanding of the role the tests were to play in the evaluation of their project. How each of these factors affected the test framework is discussed in the following sections.

National and State Science Standards

As stated in Chapter Two, Project 2061 was a long-term effort launched in 1985 by the Association for the Advancement of Science to reform science, mathematics, and technology education. The *National Science Education Standards* were based in part on Project 2061's *Benchmarks for Science Literacy*, published in 1993, that specified how students should progress toward science literacy. The National Science Foundation was one of the major funding agencies for the *Standards* project.

One of the ways the *Standards* affected the scope of the construct in phase 1 of this test development task was through its impact on the goals of the Math Science Partnership program, as identified in the NSF Program Solicitation 02-061 (Table 10).

Table 10. MSP goals and applicable *Standards*

MSP Goals	**Applicable *NSES***
Enhance schools' capacity to provide challenging curricula for all students and encourage more students to succeed in advanced courses in mathematics and the sciences.	• Science Teaching Standards • Science Content Standards • Assessment Standards
Increase the number, quality and diversity of mathematics and science teachers, especially in underserved areas;	• Professional Development Standards • Science Education System Standards
Engage and support scientists, mathematicians, and engineers at local universities and local industries to work with K-12 educators and students;	• Science Education Program Standards • Science Education System Standards
Contribute to a greater understanding of how students effectively learn mathematics and science and how teacher preparation and professional development can be improved;	• Science Teaching Standards • Professional Development Standards • Assessment Standards • Science Content Standards
Promote institutional and organizational change in education systems — from kindergarten through graduate school —to sustain partnerships' promising practices and policies.	• Science Education Program Standards • Science Education System Standards
(NSF, 2005, p. [4]	

Through this program, NSF awards competitive, merit-based grants to partnerships composed of institutions of higher education, local K-12 school systems, and their supporting partners (Figure 5). NSF Program Solicitation 02-061 was one of the first solicitations issued under the MSP program, and the TASC project was one of the grants awarded under this solicitation.

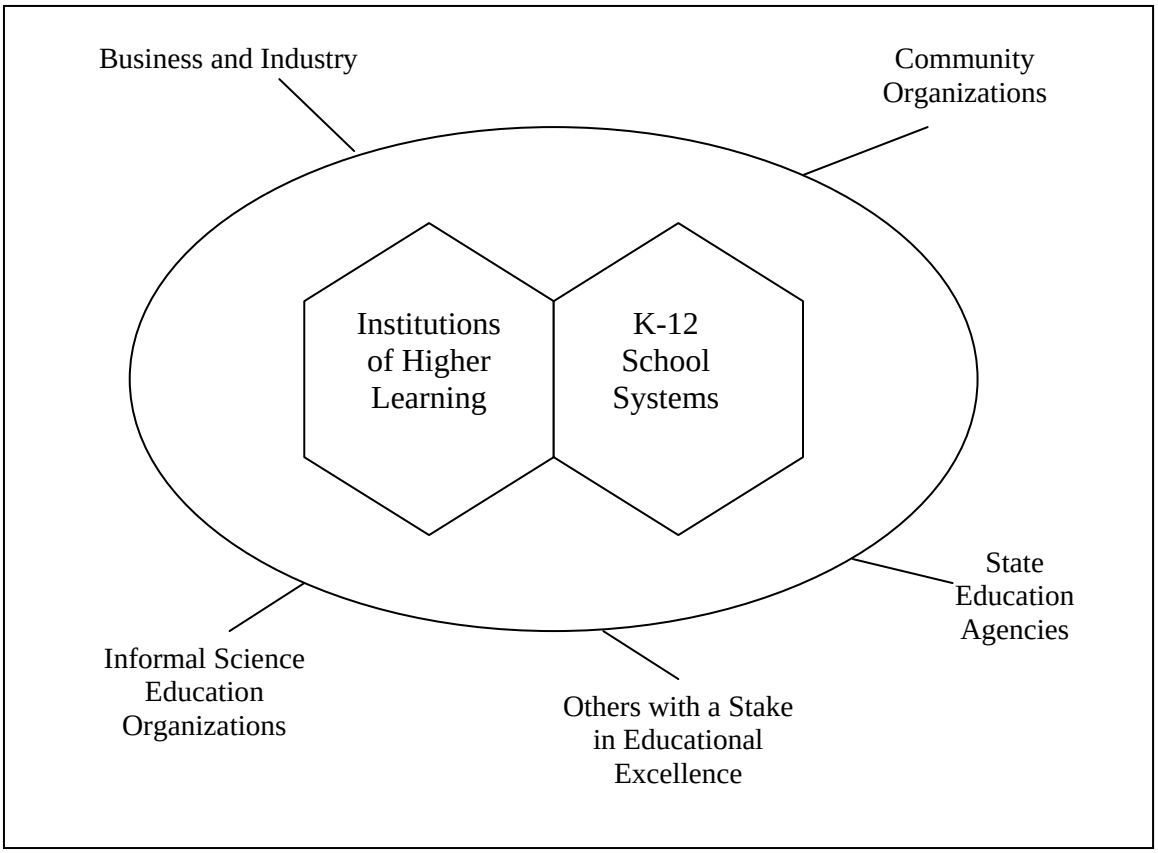


Figure 5. MSP partners (NSF, 2005, p. [2])

Another way the *Standards* affected the scope of the construct in phase 1 of this test development task was through the content of instruction. That is, the standards for science content outline what students should know, understand, and be able to do over the course of K-12 education. The North Carolina Department of Public Instruction incorporated these into the 2004 NC Standard Course of Study (NCSCS) for Science. For instance, the 2004 NCSCS for Science states that:

> The science component of the North Carolina *Standard Course of Study (SCS)* was created . . . by establishing competency goals and objectives for teaching and learning science in all grades. It contains the concepts and theories, strands, skills, and processes on which all science instruction should be based. In addition, the

> curriculum defines and illustrates the connections between the *National Science Education Standards*, the *Benchmarks for Scientific Literacy*, and the state standards. The *SCS* is a guide to stronger, more relevant science education for every student. (NC DPI, 2004, p. 6)

In addition, the 2004 NCSCS for Science states that its goal is to achieve scientific literacy, incorporating the *National Science Education Standards* definition of scientific literacy as "the knowledge and understanding of scientific concepts and processes required for scientific decision making, participation in civic and cultural affairs, and economic productivity." (p. 22). The *National Science Education Standards* elaborate further, stating:

> Scientific literacy means that a person can ask, find, or determine answers to questions derived from curiosity about everyday experiences. It means that a person has the ability to describe, explain, and predict natural phenomena. Scientific literacy entails being able to read with understanding articles about science in the popular press and to engage in social conversation about the validity of the conclusions. Scientific literacy implies that a person can identify scientific issues underlying national and local decisions and express positions that are scientifically and technologically informed. A literate citizen should be able to evaluate the quality of scientific information on the basis of its source and the methods used to generate it. Scientific literacy also implies the capacity to pose and evaluate arguments based on evidence and to apply conclusions from such arguments appropriately. (NRC, 1996, p. 22)

Again, the NCSCS for Science incorporates the *Standards'* description of scientific literacy in its "tenets of scientific literacy" that includes the ability to:

- Find or determine answers to questions derived from everyday experiences.
- Describe, explain, and predict natural phenomena.
- Understand articles about science.
- Engage in non-technical conversation about the validity of conclusions.
- Identify scientific issues underlying national and local decisions.

- Pose explanations based on evidence derived from one's own work.

(NC DPI, 2004, p. 8)

All these abilities include the *use*, rather than mere recitation, of scientific knowledge.

Tables 11 and 12 set forth the elementary and middle grade science content standards, respectively, as presented in the *National Science Education Standards* and in the 2004 NC Standard Course of Study for K-8 science. These tables clearly demonstrate how the *Standards'* unifying concepts, science as inquiry, science as technology, science in personal and social perspectives, the nature and history of science, as well as content areas are incorporated throughout the 2004 NCSCS for Science.

Table 11 Elementary grade science content standards

National Science Education Standards **Content Standards Grades K-5**		***NC Standard Course of Study for Science*** **Elementary Grades (K-5)**	
UNIFYING CONCEPTS AND PROCESSES	• Systems, order, and organization • Evidence, models, and explanation • Change, constancy, and measurement • Evolution and equilibrium • Form and function	UNIFYING CONCEPTS	• Systems, Order and Organization. • Evidence, Models, and Explanation. • Constancy, Change, and Measurement. • Evolution and Equilibrium. • Form and Function.
SCIENCE AS INQUIRY	• Abilities necessary to do scientific inquiry • Understandings about scientific inquiry	SCIENCE AS INQUIRY	• Abilities necessary to do scientific inquiry. • Abilities necessary to understand, to use, and to apply the unifying concepts and processes of science including: o evidence, explanation, measurement. o ordering, organizing. o changes (time, rate, scale, patterns, trends, cycles). o Systems. ▪ boundaries.

National Science Education Standards **Content Standards Grades K-5**		*NC Standard Course of Study for Science* **Elementary Grades (K-5)**	
			▪ components. ▪ resources. ▪ flow. ▪ feedback. o form, function, equilibrium. o models.
PHYSICAL SCIENCE	• Properties of objects and materials • Position and motion of objects • Light, heat, electricity, and magnetism		• Grade 5 --Goal 4: forces and motion
LIFE SCIENCE	• Characteristics of organisms • Life cycles of organisms • Organisms and environments		• Grade 3 --Goal 1: plant growth and adaptations --Goal 4: skeletal and muscles systems of the human body • Grade 5 --Goal 1: plants and animals
EARTH AND SPACE SCIENCE	• Properties of earth materials • Objects in the sky • Changes in earth and sky		• Grade 3 --Goal 2: soil properties --Goal 3: earth/moon/sun system • Grade 5 --Goal 2: landforms --Goal 3: weather and climate
SCIENCE AND TECHNOLOGY	• Abilities of technological design • Understandings about science and technology • Abilities to distinguish between natural objects and objects made by humans	SCIENCE AND TECHNOLOGY	• Ability to use and create technological designs. • Understanding about technology and design. • Ability to distinguish between natural and human made objects.
SCIENCE IN PERSONAL AND SOCIAL PERSPECTIVES	• Personal health • Characteristics and changes in populations • Types of resources	SCIENCE IN PERSONAL AND SOCIAL PERSPECTIVES	• Impacts of science and technology on their daily lives. • The relationship of science

National Science Education Standards **Content Standards Grades K-5**		*NC Standard Course of Study for Science* **Elementary Grades (K-5)**	
	• Changes in environments • Science and technology in local challenges		to personal health and welfare. • Characteristics of and changes in populations. • Applications of science and technology to local challenges.
HISTORY AND NATURE OF SCIENCE	• Science as a human endeavor	NATURE OF SCIENCE	• Science as a human endeavor • Science as inquiry • The nature of scientific inquiry
(NRC, 1996, p. 109)		(NC DPI, 2004, pp. 24-25)	

Table 12. Middle grades science content standards

National Science Education Standards **Content Standards Grades 5-8**		*NC Standard Course of Study for Science* **Middle Grades (6-8)**	
UNIFYING CONCEPTS AND PROCESSES	• Systems, order, and organization • Evidence, models, and explanation • Change, constancy, and measurement • Evolution and equilibrium • Form and function	UNIFYING CONCEPTS	• Systems, Order and Organization. • Evidence, Models, and Explanation. • Constancy, Change, and Measurement. • Evolution and Equilibrium. • Form and Function.
SCIENCE AS INQUIRY	• Abilities necessary to do scientific inquiry • Understandings about scientific inquiry	SCIENCE AS INQUIRY	• Ability to do scientific inquiry. • Understanding about scientific inquiry. • Ability to perform safe and appropriate manipulation of materials, scientific equipment, and technology. • Mastery of integrated process skills. -acquiring, processing, and interpreting data -identifying variables and their relationships -designing investigations

National Science Education Standards **Content Standards Grades 5-8**		*NC Standard Course of Study for Science* **Middle Grades (6-8)**	
			-experimenting -analyzing investigations -constructing hypotheses -formulating models
PHYSICAL SCIENCE	• Properties and changes of properties in matter • Motions and forces • Transfer of energy	PHYSICAL SCIENCE	• Grade 8 --Goal 3: hydrosphere --Goal 4: chemistry
LIFE SCIENCE	• Structure and function in living systems • Reproduction and heredity • Regulation and behavior • Populations and ecosystems • Diversity and adaptations of organisms	LIFE SCIENCE	• Grade 8 --Goal 6: cell theory --Goal 7: microbiology
EARTH AND SPACE SCIENCE	• Structure of the earth system • Earth's history • Earth in the solar system	EARTH AND SPACE SCIENCE	• Goal 8 --Goal 5: evolution of organisms and landforms
SCIENCE AND TECHNOLOGY	• Abilities of technological design • Understanding about science and technology	SCIENCE AND TECHNOLOGY	• What technologies are. • Ability to perform technological design. • Understanding science and technology.
SCIENCE IN PERSONAL AND SOCIAL PERSPECTIVES	• Personal health • Populations, resources, and environments • Natural hazards • Risks and benefits • Science and technology in society	SCIENCE IN SOCIAL AND PERSONAL PERSPECTIVES	• Personal and community health. • Population dynamics. • Environmental quality. • Natural and human-induced hazards. • Science and technology in local, national, and global challenges. • Careers in science and technology.

National Science Education Standards **Content Standards Grades 5-8**		*NC Standard Course of Study for Science* **Middle Grades (6-8)**	
HISTORY AND NATURE OF SCIENCE	• Science as a human endeavor • Nature of science • History of science	NATURE OF SCIENCE	• Science as a human endeavor • Nature of scientific knowledge Historical perspectives
(NRC, 1996, p. 110)		(NC DPI, 2004, pp. 52-53)	

TASC, in turn, stated that their curriculum units were based on the NC Standard Course of Study for Science. Table 13 provides a brief description of each unit and the NC SCS competency goal with which the unit was aligned as stated on the TASC website.

Table 13. TASC science curriculum units

TASC Curriculum Unit	**Description**	**NCSCOS Goal**
Gr 3 Human Body	Engage students in thoughtful activities about the form and function of a most remarkable machine, their own bodies. Students build mechanical models to demonstrate how muscles power human movement and develop an appreciation for the design and coordination of the human body. Publisher: FOSS (Full Option Science System), Delta Education	NCSCOS Goal 4: The learner will conduct investigations and use appropriate technology to build an understanding of the form and function of the skeletal and muscle systems of the human body.
Gr 3 Investigating Objects in the Sky	Students explore and describe the position, appearance, and motion (or apparent motion) of objects in the sky, specifically the Moon, the Sun, and the stars. They use their shadows to determine the changing position of the Sun in the daytime sky and use direct observations to describe the changing position of the Moon during the day and at night and of the stars in the nighttime sky. Students also observe that the Moon appears to change its shape every day in a repeating pattern that takes approximately one month. Publisher: TRACS (Teaching Relevant Activities for Concepts & Skills), BSCS	NCSCOS Goal 3: The learner will make observations and use appropriate technology to build an understanding of the earth/moon/sun system.
Gr 3 Plant Growth & Development	Students observe each stage in the life cycle of a simple plant. Students plant seeds and watch the	NCSCOS Goal 1: The learner will conduct

TASC Curriculum Unit	Description	NCSCOS Goal
	seedlings emerge. They thin and transplant seedlings. As they watch plants grow, students learn that plants need nutrients from the soil, as well as water and light, to thrive. To explore the interdependence of living things, students pollinate the flowers with dried honeybees. Finally, they harvest mature seeds and determine seed yields. Publisher: STC (Science and Technology for Children), Carolina Biological Supply Company	investigations and build an understanding of plant growth and adaptations.
Gr 3 Soils	Examinations of properties of different soil components. Students characterize the various soil components, then use this information to identify mystery soils and analyze characteristics of their local soils. Publisher: STC (Science and Technology for Children), Carolina Biological Supply Company	NCSCOS Goal 2: The learner will conduct investigations to build an understanding of soil properties.
Gr 5 Ecosystem	Students begin the unit by setting up a terrarium in which they grow grass, mustard, and alfalfa plants. They then add crickets and isopods. They also set up an aquarium into which they introduce snails, guppies, elodea, algae, and duckweed. By connecting the terrarium and aquarium bottles to create an "ecocolumn," students are able to observe the relationship between the two environments and the organisms living within them. Using test ecocolumns that contain only plants, students simulate the effects of pollutants—such as road salt, fertilizer, and acid rain—on an environment. Students then use a food chain wheel to make inferences about the effects these pollutants might have on their own miniature ecosystems. Later, students read about, explore, and discuss the Chesapeake Bay as a model ecosystem. They analyze this ecosystem from the viewpoint of various users—-waterman, dairy farmer, land developer, recreational boater, and resident—-and present their findings to the class. This activity enables students to appreciate the trade-offs that must be made to reach mutually acceptable solutions to environmental problems. Publisher: STC (Science and Technology for Children), Carolina Biological Supply Company	NCSCOS Goal 1: The learner will conduct investigations to build an understanding of the interdependence of plants and animals.
Gr 5 Investigating Weather Systems	A variety of explorations of weather systems. Students discover the major factors that affect weather, including latitude, altitude, and proximity to bodies of water. They make physical models that	NCSCOS Goal 3: The learner will conduct investigations and use appropriate technology to

TASC Curriculum Unit	Description	NCSCOS Goal
	illustrate the driving forces of weather. They keep records of weather changes outside their classroom and graph the resulting data. Publisher: TRACS (Teaching Relevant Activities for Concepts & Skills), BSCS	build an understanding of weather and climate.
Gr 5 Landforms	This unit consists of five investigations that introduce students to these fundamental concepts in earth science: change takes place when things interact; all things change over time; patterns of interaction and change are useful in explaining landforms. Students also learn about some of the tools and techniques used by cartographers and use them to depict landforms. Publisher: FOSS (Full Option Science System), Delta Education	NCSCOS Goal 2: The learner will make observations and conduct investigations to build an understanding of landforms.
Gr 5 Motion & Design	Investigations of motion of vehicles and challenges in technological design and engineering. Students create vehicles and use them to explore the effects of force, friction, and wind resistance on speed and distance. They graph data gathered about the motion of their vehicles under various forms of power. They are challenged to build their own vehicles to meet specifications such as distance traveled in a given time and cost. Publisher: STC (Science and Technology for Children), Carolina Biological Supply Company	NCSCOS Goal 4: The learner will conduct investigations and use appropriate technologies to build an understanding of forces and motion in technological designs.
Gr 8 Earth History	Students investigate sedimentary rocks and fossils from the Grand Canyon to discover clues that reveal Earth's history. They consider the processes that created them and compare evidence discovered in the rocks to present-day geologic processes and contemporary life forms. Then students use these data to make inferences about past organisms, environments, and events that occurred on Earth over its history. Publisher: FOSS (Full Option Science System), Delta Education	• NCSCOS Goal 1: The learner will design and conduct investigations to demonstrate an understanding of scientific inquiry. • NCSCOS Goal 2: The learner will develop demonstrate an understanding of technological design. • NCSCOS Goal 5: The learner will conduct investigations and utilize appropriate technologies and information systems to build an understanding of evidence of evolution in organisms and

TASC Curriculum Unit	Description	NCSCOS Goal
		landforms.
Gr 8 MicroLife	Students study microbiology; cell size, structure, function and permeability, and systems of classification. They explore the function of the immune system and the growth of antibiotic-resistant organisms. A project on disease develops research skills. Publisher: SEPUP, Lawrence Hall of Science, Lab-Aids, Inc.	• NCSCOS GOAL 6: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of cell theory. • NCSCOS GOAL 7: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of microbiology.
Retrieved 4/6/2005 from http://tasc.pratt.duke.edu/index.php		

In summary, the purpose of the test and scope of the construct was affected by national and state science standards. The *National Science Education Standards*, that were based upon Project 2061's *Benchmarks for Science Literacy*, were reflected in the MSP goals as set forth in NSF Program Solicitation 02-061. TASC was one of the MSP projects awarded under this particular solicitation. TASC provided professional development training to participating teachers on science units that were aligned with grade-specific science competency goals of the 2004 NC Standard Course of Study. The NC SCS science competency goals were, in turn, written to be more aligned with the *National Science Education Standards*.

NSF's Definition of "Evidence" in a Project Evaluation

In the NSF's *2002 User-Friendly Handbook for Project Evaluation*, the authors acknowledge that evaluations "are designed for various audiences, including funding agencies, policymakers in governmental and private agencies, project staff and clients, researchers in academic and applied settings, and various other stakeholders". (p. 45) In their discussion of qualitative information versus quantitative information, the authors state:

> The major stakeholders for NSF projects are policymakers within NSF and the federal government, state and local officials, and decisionmakers in the educational community where the project is located. In most cases, decisionmakers at the national level tend to favor quantitative information because these policymakers are accustomed to basing funding decisions on numbers and statistical indicators. On the other hand, many stakeholders in the educational community are often skeptical about statistics … and consider the richer data obtained through qualitative research to be more trustworthy and informative. A particular case in point is the use of traditional test results, a favorite outcome criterion for policymakers, school boards, and parents, but one that teachers and school administrators tend to discount as a poor tool for assessing true student learning. (p. 45; emphasis added.)

In their discussion of the use of tests, the authors state that tests are appropriate to use when

> one wants to gather information on the status of knowledge or the change in status of knowledge over time. … Changes in test performance are frequently used to determine whether a project has been successful in transmitting information in specific areas or influencing the thinking skills of participants. (p. 56; emphasis added.)

The NSF Solicitation 02-061 made it clear that all proposed partnerships would establish "results-oriented, accountable projects that implement evidence-based educational practices resulting in improved preK-12 student outcomes" and generate and sustain "an exceptionally competent mathematics and science preK-12 teaching workforce". Partnerships were directed to provide data that included both student and teacher indicators in (math and) science. (NSF, 2002a, p. 5-6).

To plan their evaluations, proposed partnerships were directed by the solicitation to:

> carefully plan project evaluation to guide the annual assessment of progress and to measure the impact of the effort. This section should include the means by which the partners document, measure, and report on the resources, allocations, programs, policies, procedures, and measurable outcomes as they bear on accountability for science and mathematics improvement related to the MSP program goals. In the formative sense, evaluation should provide evidence of the strengths and weaknesses of the effort being implemented, facilitating the partnership's understanding of what works. The evaluation should also be designed to respond to the summative need of analyzing both qualitative and quantitative data to determine the effectiveness of the partnership in contributing to positive institutional changes and student academic outcomes.
> (p. 13; emphasis added)

Lastly, the solicitation provided the review criteria by which proposals would be judged. Particularly relevant to this study was the "Data" criterion outlined in the solicitation:

> … Reviewers will consider the following types of questions and apply them to reviews of proposals for comprehensive or targeted awards, as appropriate.
>
> . . .

DATA

- Will data collection activities provide disaggregated data by race/ethnicity, gender, socioeconomic status, and disability, as well as valid indices of student performance in mathematics and science?
- Are annual benchmarks of progress related to programmatic goals (for both preK-12 students and teachers) and strategic actions indicating both short-term and long-term outcomes for all partners established against a baseline?
- Are the benchmarks reasonable and appropriate in demonstrating an anticipated rate of improvement that exceeds that of locals in which no MSP investment exists?
- Is there evidence of effective mathematics and science assessment systems to be utilized in order to gather, interpret, and use reliable student achievement data that can be used to inform the MSP planning and decision-making process?
- To what extent does the accountability system encompass the appropriate use of data, including the tracking of students' outcomes (e.g., performance, attitudes, and enrollment in high school STEM advanced courses)?

(p. 17; emphasis added)

In summary, federal decision-makers prefer quantitative data on which to base their funding decisions. The overall objective of NSF's MSP program is to improve student outcomes in (math and) science by all students, at all preK-12 levels. Through its Solicitation 02-061, NSF requested that proposed partnerships between higher education and local school districts establish "results-oriented, accountable projects that implement evidence-based educational practices resulting in improved preK-12 student outcomes" (p. 5) and that data collection include teacher and student indictors in (math and) science. The proposed evaluation plans of the partnerships were to include "both qualitative and quantitative data to determine the effectiveness of the partnership in contributing to positive institutional changes and student academic outcomes" (p. 13) in order to provide evidence that would be used by NSF to inform the MSP planning and decision-making process.

Whereas NSF cites the use of qualitative and quantitative data sources as appropriate evidence of effectiveness in project evaluations, it appears that for funding purposes, they prefer quantitative data. Thus, when a MSP project proposes improvement in science content knowledge, the "strongest"—most highly favored by NSF—evidence it could provide would be student and teacher test data because NSF frequently uses changes in test performance to determine whether a project has been successful in accomplishing its goals and objectives.

The MSP Project's Understanding of the Role of the Tests in Its Project Evaluation

As mentioned previously, the stated purpose of a test is foundational to the test development process. In response to the NSF Program Solicitation 02-061, TASC proposed the following:

> An assessment/accountability system for science K-8 is non-existent in North Carolina at this time. However, TASC will help to put such a system in place and improve its quality. The NC Department of Public Instruction (DPI) is the TASC partner contracted to assess student performance. DPI's TASC assessment will contribute to the 2007 statewide assessment, and it will of necessity assess science process. The involvement of the Department of Public Instruction in TASC should help steer development of the state toward authentic assessment of student performance in science.
> (TASC (2002a), p. 2; emphasis added).

In addition to creating an assessment/accountability system for K-8 science emphasizing authentic assessment of student performance in science, TASC's proposal included the following two (of seven) objectives:

> Improve science teaching in participating schools by bolstering teacher content knowledge, improving the use of inquiry-based curriculum, and improving inquiry-based teaching
>
> Prepare teacher leaders and "curriculum unit experts" to help and train their colleagues.
> (TASC (2002a), p. 8).

TASC proposed to assess these objectives using student measures and teacher measures. Student measures were to include:

> … curriculum units [that] will contain a small set of knowledge level test items and science process test items (4-5 items total), tied to the NC Standard Course of Study and selected such that they are also covered in non-participating classrooms. Science process skills tested would match problem solving tasks analogous to those in curriculum units but would also be familiar to students of non-participating teachers. Again, the annual milestone is met when a statistically significant positive difference is observed [between students of participating and non-participating teachers in partner districts].
> (TASC (2002a), p. 9).

TASC proposed to measure the changes in teachers' knowledge by "performance on teacher post-training science content tests". (p. 9)

Under the assessment/accountability section of TASC's proposal, an elementary science consultant for the NC Department of Public Instruction (DPI) was to coordinate the evaluation of the project, working with an external evaluator. The DPI consultant was "to provide expertise with the development and delivery of the science content and process skills tests for grades 3, 5, and 8." In addition:

> [a]n NC-ISE [NC Infrastructure for Science Education, a NC DPI program] committee of university science educators and student assessment specialists will guide the selection of items for development of the content and process skills test.

> They will also assist with the reliability and validity measures of these tests and the random distribution of the tests. Assessment of student learning in an inquiry-based science unit is a major focus of this project; therefore, the project staff will work closely with the NC-ISE evaluation committee to develop student assessment measures that will document student conceptual development and growth in science and mastery of knowledge and science process skills. Program assessment and evaluation will address the following:
>
> > 2) Student Impact. … standardized science tests will be developed for the third, fifth, and eighth grade levels, piloted, and administered to random paired groups of students from participating and non-participating classrooms. These tests will measure students' understanding of science process skills and relevant content for that grade level. Rubrics for evaluating student responses to open-ended problems will be developed and field-tested. …
> >
> > 3) Teacher impact. Teacher understanding of inquiry-based instruction, science process skills and relevant content concepts will be evaluated through interviews and standardized surveys (Weiss, et al., 2001). …
>
> (TASC (2002a), p. 15).

Using tables, the TASC evaluation plan indicated that surveys, interviews, and observations would be used to determine the project's impact on teachers' material use, instructional skills, science attitudes, and content knowledge. To measure teachers' content knowledge, the evaluation plan document stated:

> Short curriculum unit topic content tests will be developed comprised of a small set of knowledge level and problem-solving test items tied to the NC Standard Course of Study and the curriculum topics for grades K-8. The project staff will develop these tests. Participating scientists will review these tests for content validity and reliability measurements will be completed as well.
> (TASC (2002b), p. 1).

In addition, the TASC evaluation plan indicated that student test, survey, and interview data would be used to determine the project's impact on students' science

content knowledge and science process skill. To measure student content knowledge and process skill, the evaluation plan stated:

> Short curriculum unit topic content tests will be developed comprised of a small set of knowledge level and problem-solving test items tied to the NC Standard Course of Study and the curriculum topics for grades K-8. The project staff will develop these test, which will be the same as the tests given to teachers …. Participating scientists will review these tests for content validity and reliability measurements will be completed as well. The tests will be given to the students at the end of each curriculum unit to determine demonstration of content knowledge.
>
> An assessment committee will develop student tests to be given at grades 3, 5 and 8. These tests will measure student science content knowledge and process skills. Under the supervision of [name] of the NC DPI, this committee will be comprised of university science educators and teachers. The committee will guide the selection of the test items and will assist with the reliability and validity measures of these tests. These pilot tests will be given to students in participating school systems.
>
> (TASC (2002b), pp. 5-6)

TASC, in its proposal to NSF, communicated its multidimensional understanding of the role of the to-be-developed tests. That is, two sets of tests were to be developed:

- short curriculum unit topic content tests:
    1. These tests were to be made up of a small set of knowledge level and problem-solving test items tied to the NC Standard Course of Study and the curriculum topics for grades K-8.
    2. The tests were to be developed by the project staff, who would also review the tests for content validity and complete reliability measurements.
    3. Teachers and students were to receive the same test.
    4. Both teacher and student test results would be used to evaluate changes in teacher and student science content knowledge.

- large assessment state tests:
    1. These tests were to be part of a state-wide assessment/accountability system.
    2. These tests were to be designed to authentically measure student performance in science.
    3. Item selection for these content and process skills tests was to be guided by an NC-ISE committee of university science educators and student assessment specialists (i.e., NC DPI), who also were to "assist with the reliability and validity measures of these tests" as well as the random distribution of the tests. In addition, the project staff were to were to work closely with the NC-ISE evaluation committee in the development of these student assessment measures that were intended to document student conceptual development and growth in science and mastery of knowledge and science process skills.
    4. Standardized science tests were to be developed for the third, fifth, and eighth grade levels, piloted, and administered to random paired groups of students from participating and non-participating classrooms.
    5. Student test data were to be used to evaluate whether the project met its annual milestone; that is, the observation of a statistically significant positive difference between students of participating and non-participating teachers in partner districts.

Test Specifications (Phase 2)

Once the test purpose and the construct have been established, the next phase of test development includes the development and evaluation of the test specifications. This phase includes such tasks as determining:

- the format of items, tasks, or questions;
- the response format or conditions for responding;
- the type of scoring procedures;
- time restrictions, if applicable;
- test length;
- test blueprint:
    - instructional objectives to be measured and
    - cognitive skills to be required of examinees;
- characteristics of intended test-takers; and
- procedures for administration.

In this study, TASC established the purpose and construct of the to-be-developed tests (phase 1 tasks). CERE, in its subcontract with TASC, was to construct and evaluate the pilot tests (phase 3 tasks) and to assemble and evaluate the operational tests (phase 4 tasks). A logical expectation would be that TASC would have established the tests' specifications (phase 2 tasks) prior to CERE beginning the test construction tasks. However, this was not the case; rather than working on test specifications, TASC initiated work on item development, tasks that are part of phase three of test

development. Even so, the next section discusses the factors that affected phase 2 of the test development process.

Factors that Affected Phase 2 (Test Specifications)

Factors that affected this phase of test development included national and state science standards, the TASC project personnel's understanding of the *role* of the to-be-developed tests in their project evaluation, and TASC's understanding of the *process* of test development. The following subsections discuss how each of these factors affected the test specifications phase of the test development process.

National and State Science Standards

As stated in the previous section (Test Framework), scientific literacy for all is a goal promulgated by the *National Science Education Standards*, and, in turn, by the 2004 NC Standard Course of Study in Science. The *Standards* are based on the premise that "science is an active process". Thus, not only must students have "hands on" experiences but they must have "minds on" experiences as well. The *Standards* state: "Inquiry is central to science learning." (NRC, 1996, p. 2).

Likewise, the 2004 NC Standard Course of Study states:

> The revised North Carolina Standard Course of Study takes students beyond science as merely a body of knowledge to science as inquiry. It requires students to combine science and scientific knowledge with scientific reasoning and critical thinking. Engaging students in scientific inquiry helps them develop:
> - An understanding of scientific concepts.
> - An appreciation of how we know what we know in science.
> - An understanding of the nature of science, along with the skills to become independent discoverers of the natural world.
> - The disposition to use the skills and attitudes associated with science.

> Science as inquiry is key to organizing and guiding students' activities. Students in all grades and in every scientific discipline should have the opportunity to ask questions, plan and conduct investigations, use appropriate tools and techniques to gather data, think critically and logically about relationships between evidence and explanations, and communicate arguments.
> (NC DPI, 2004, p. 17; emphasis added.)

TASC's purpose was to provide North Carolina K-8 students with opportunities to learn to think as scientists, that is, critically, creatively, and independently. To accomplish this purpose, TASC provided teacher trainings that focused on science content, inquiry-based teaching, and effective use of science materials. In this way, TASC—using NSF-approved science curriculum kits—trained K-8 teachers in participating school districts to create situations in which students would take the role of scientists. That is, students would observe and question phenomena, pose explanations of what they would see, devise and conduct tests to support or contradict their theories, analyze data, draw conclusions from experimental data, design and build models, and discuss their findings. (Retrieved May 9, 2005, from http://tasc.pratt.duke.edu/about.overview.php).

This inquiry-based focus of science instruction affected the tests' blueprints. One of the outcomes of NC science education is that students learn to think critically and creatively, and the NC Standard Course of Study provides the competencies that students should demonstrate. As previously discussed in Chapter Three, the tables of specifications developed in the context of this study each consisted of a two-dimensional grid. One dimension included the instructional objectives (applicable to the curriculum unit taught by the TASC trainers) derived from 2004 NC Standard Course of Study for

Science. The second dimension included the cognitive processes derived from the NC Thinking Skills, the taxonomy adopted by the NC Department of Public Instruction (one of the partners in the TASC project). Table 14 illustrates the incorporation of this two-dimensional grid into one of the test blueprints.

Table 14. Grade 3 *Plant Growth & Development* test blueprint

Grade 3 Competency Goal 1	**Objectives**	**% of time spent on objective**	**Minimum number of items to be created**	**NC Thinking Skills**					
				Organizing	**Applying**	**Analyzing**	**Generating**	**Integrating**	**Evaluating**
The learner will conduct investigations and build an understanding of plant growth and adaptations.	1.01 Observe and measure how the quantities and qualities of nutrients, light, and water in the environment affect plant growth. 1.02 Observe and describe how environmental conditions determine how well plants survive and grow in a particular environment. 1.03 Investigate and describe how plants pass through distinct stages in their life cycle including. • Growth. • Survival. • Reproduction 1.04 Explain why the number of seeds a plant produces depends on variables such as light, water, nutrients, and pollination.								

Grade 3 Competency Goal 1	Objectives	% of time spent on objective	Minimum number of items to be created	NC Thinking Skills					
				Organizing	Applying	Analyzing	Generating	Integrating	Evaluating
	1.05 Observe and discuss how bees pollinate flowers. 1.06 Observe, describe and record properties of germinating seeds.								

On March 22, 2005, the CERE acting director and this researcher met with the TASC project director to discuss the CERE-TASC scope of work (in the unsigned TASC-CERE subcontract). Because one of the objectives of NC science instruction was to equip students to think critically, the TASC project director indicated that the items were to target higher level thinking skills (i.e., no knowledge-level questions) and were to be of "medium to hard" difficulty. This is reflected in Table 14 (above) with the omission of the knowledge-level NC Thinking Skill in the sample test blueprint.

In summary, national and state science standards emphasized the active nature of science instruction. This inquiry-based science learning was foundational to TASC's teacher training. The test blueprints were affected in that items to be written for the tests were to reflect both NC SCS instructional objectives *and* NC thinking skills, were to require higher level thinking skills (i.e., no knowledge-level items), and were to include items of "medium to hard" difficulty.

The MSP Project's Understanding of the Role of the Tests in Its Project Evaluation

TASC's initial understanding of the role of the to-be-developed tests in its project evaluation was that test results would be used to provide quantitative evidence to the funding agency that TASC was meeting the teacher and student outcome goals regarding increased science content knowledge. As stated previously, TASC, in its 2002 proposal to NSF, proposed the development of *two* sets of tests:

- Short curriculum unit topic content tests that would be:
    - made up of a small set (four or five) of *knowledge level* and *problem-solving items* tied to the NC Standard Course of Study and the curriculum topics for grades K-8;
    - developed by the "project staff", who would review the tests for content validity and evaluate test reliability;
    - administered *to* TASC-participating teachers; and
    - administered *by* TASC-participating *and* non-TASC participating teachers to their students.
- Large assessment state tests that would be:
    - part of a state-wide assessment/accountability system, and
    - made up of items designed to authentically measure student performance in science with item selection guided by a collaboration of the partnership's university science educators and student assessment specialists.

According to its evaluation plan (TASC, 2002b), one of the ways TASC expected to impact teachers was through teachers' increased science content knowledge. The data source for this particular outcome was a “pre-post science content knowledge test for each curriculum unit”. Thus, TASC expected to use the test results to provide quantitative evidence to the funding agency that of the teachers participating in TASC, “70% of tested teachers will demonstrate a 20% gain in science content knowledge” (TASC (2002b), p. 1).

TASC's evaluation plan also stated that one of the student outcomes was that students of participating teachers would demonstrate knowledge of science content and skill with science process. The data source—“curriculum unit tests by teacher”—was to be the same curriculum units tests administered to the TASC-participating teachers. Again, TASC expected to use the test results to provide quantitative evidence to its funding agency that of the students of TASC-participating teachers, “80% of tested students will score 70% or higher on the given curriculum test” (TASC (2002b), p. 5).

A second data source for this student outcome was to be the “pilot state administered tests". According to the evaluation plan, these tests would be administered to "students of participating teachers in this program and students of non-participating teachers" and to (undefined) “treatment groups of students” (TASC (2002b), p. 5). TASC expected to use the test results to provide quantitative evidence to its funding agency that students of TASC-participating teachers would score “level 3 or above on state-administered science content and process skill test” and that there would be a "20% expected difference between treatment groups of students” (TASC (2002b), p. 5).

In its proposal, TASC clearly addressed NSF funders' preference for quantitative data by proposing to use results from yet-to-be-developed science tests as quantitative evidence that TASC was meeting its teacher and student outcome goals of increased science content knowledge.

This initial understanding of the role of the to-be-developed tests was further elaborated upon in TASC's Strategic Plan document (TASC, 2002c). In the "Year 1 Benchmarks to Measure Progress" section, TASC included student achievement benchmarks that stated:

> . . . For Year 1, TASC will take the following statistically significant positive differences between students of participating and non-participating teachers as benchmarks:
>
> - curriculum units will contain 4-5 science knowledge and process test items tied to the NC Standard Course of Study and covered in curriculum units. Tested knowledge and skills are familiar to students of non-participating teachers. The same items will be administered in participating and non-participating classrooms.
>
> (TASC (2002c), p. 1; emphasis added.)

In addition, TASC included science teacher benchmarks that stated:

> … For Year 1, TASC will take as benchmarks for improvements in … quality … of K-8 science teachers statistically significant positive changes in teachers' knowledge … as measured by:
>
> . . .
>
> - pre- and post-test scores on content and science process test items related to the curriculum units being taught, among teachers attending training sessions (first cycle).
>
> (TASC (2002c), p. 2; emphasis added.)

Thus, in addition to students of TASC-participating teachers scoring a "level 3 or above" on the yet-to-be-developed state science tests, TASC "will take" as evidence of achieving this student outcome "statistically significant positive differences" between these students and students of non-TASC-participating teachers on the shorter curriculum unit tests. Likewise, TASC "will take" as evidence of achieving their teacher outcome "statistically significant positive differences" in TASC-participating teachers' pre- and post-test scores on the short curriculum unit tests.

TASC's initial understanding of the role of the to-be-developed tests was reiterated in its year two scope of work statement for NC DPI that stated:

> This Agreement spans the period from January 1, 2004 to December 31, 2004. The following benchmarks for participating teachers and their students are listed in the TASC strategic plan. The DPI Instruction Evaluation & Assessment Team is subcontracted by TASC (Duke University) to measure progress toward these benchmarks:
>
> *Benchmarks for Improvement in Number, Quality, and Diversity of K-8 Science Teachers*
> For these benchmarks, TASC will take as evidence statistically significant positive changes in teachers' knowledge … as measured by:
> . . .
> - pre- and post-test scores on content and science process test items related to the curriculum units being taught, among teachers attending training sessions.
>
> *Student Achievement Benchmarks*
> TASC will take as benchmarks statistically significant positive differences between students of participating and non-participating teachers in partner districts. These include:
> . . .
> - curriculum units will contain 4-5 knowledge level and science process test items tied to the NCSCS and covered in non-participating classrooms. The annual benchmark would be met by statistically significant positive differences between students of participating and non-participating teachers.

(TASC/Duke University-NC DPI year two subcontract scope of work; emphasis added.)

In this way, TASC articulated its initial understanding of the role of the to-be-developed tests in its proposed project evaluation, in its strategic (implementation) plan document, and in its subcontract with NC DPI. That is, that TASC expected to use test results as quantitative evidence to the funding agency that TASC was meeting its stated teacher and student outcome goals.

However, by early 2005, TASC's initial understanding of the role of the to-be-developed tests appeared to have changed somewhat. As stated previously in the Test Framework section, TASC subcontracted CERE/UNCG to develop tests that TASC would use "to measure improvements in content knowledge of participating teachers and their students". So it appears that TASC's stated *purpose* of the to-be-developed tests had not changed; that is, to measure changes in science content knowledge. However, in a February 14, 2005 email to the CERE acting director, the TASC project director indicated that "the question we are trying to answer" was "How does preparation of trainers affect the quality of workshops?" From this email and from the TASC logic model provided to CERE (see Figure 6 below), it appeared that TASC planned to use the test results as "evidence from training" rather than as "evidence toward … [teacher and student] benchmarks" as it indicated in its subcontract with NC DPI.

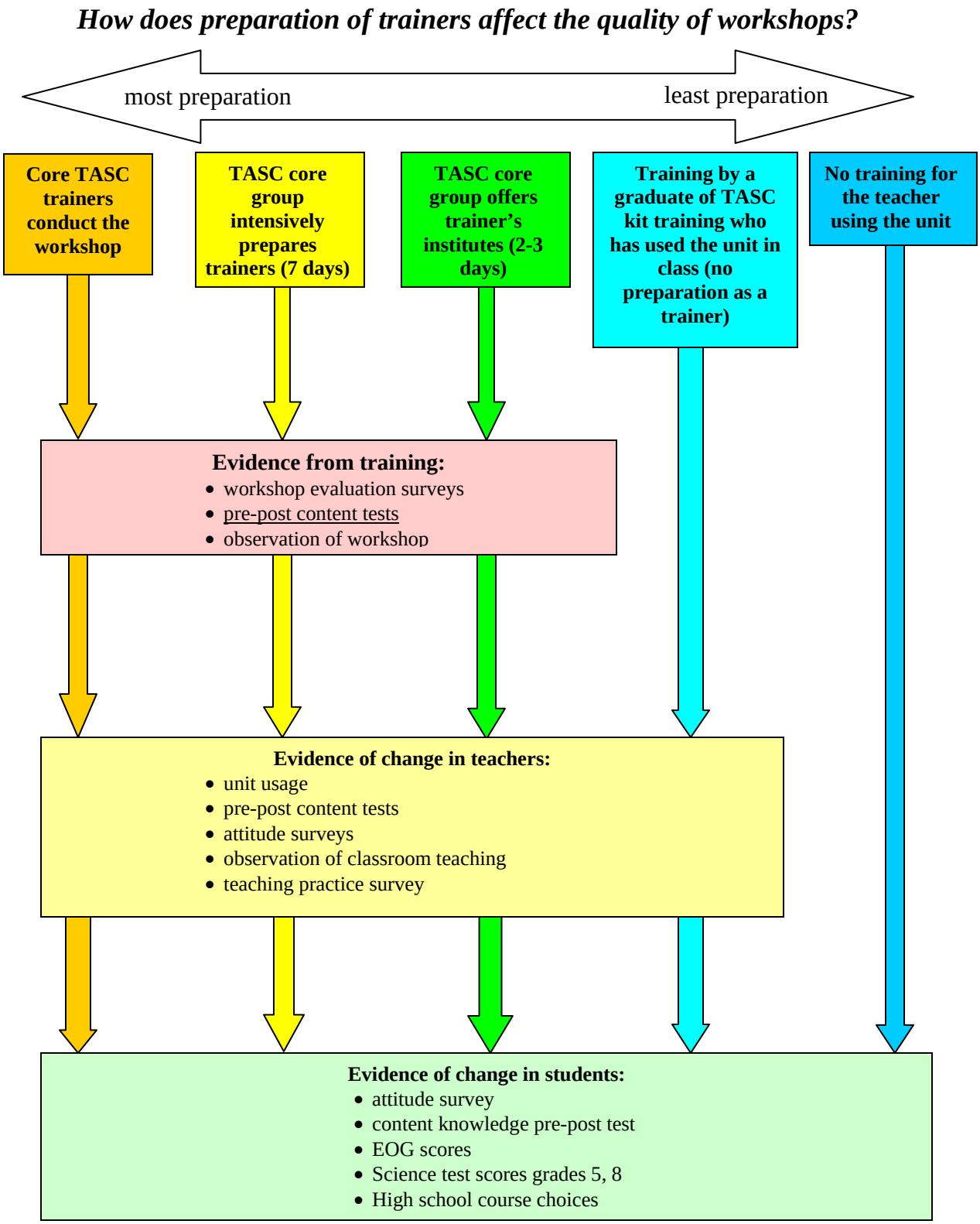


Figure 6. TASC Logic Model

The MSP Project's Understanding of the Test Development Process

TASC's communicated its understanding of the test development process in its evaluation plan, in its implementation plans, in its strategic plan, and in its subcontract with NC DPI.

Under its teacher outcome, TASC's evaluation document stated:

> Short curriculum unit topic content tests will be developed comprised of a small set of knowledge level and problem-solving test items tied to the NC Standard Course of Study and the curriculum topics for grades K-8. The project staff will develop these tests. Participating scientists will review these tests for content validity and reliability measurements will be completed as well.
>
> These tests will be administered at the onset of the teacher workshop on the curriculum unit and at the completion of the workshop. Differences in scores will be determined between the pre- and post test, expressed by percentage gain. (TASC (2002b), pp. 1-2; emphasis added.)

Under its student outcome, TASC's evaluation plan stated:

> Short curriculum unit topic content tests will be developed comprised of a small set of knowledge level and problem-solving test items tied to the NC Standard Course of Study and the curriculum unit topics for grades K-8. The project staff will develop these tests, which will be the same as the tests given to teachers as explained under Teacher Outcome #1 above [i.e., increased science content knowledge]. Participating scientists will review these tests for content validity, and reliability measurements will be completed as well. The tests will be given to the students at the end of each curriculum unit to determine demonstration of content knowledge.
>
> An assessment committee will develop student tests to be given at grades 3, 5 and 8. These tests will measure student science content knowledge and process skills. Under the supervision of … NC DPI, this committee will be comprised of university science educators and teachers. The committee will guide the selection of the test items and will assist with the reliability and validity measures of these tests. These pilot tests will be given to students in participating school systems. (TASC (2002b), pp. 5-6; emphasis added).

TASC appeared to understand that the to-be-developed tests should be content valid, be reliable, and have items tied to (aligned with) the NC SCS. TASC also appeared to understand that content experts ("university science educators and teachers") and measurement experts (NC DPI) needed to be actively involved in the test development process. Additionally, TASC appeared to understand that tests needed to be piloted before used operationally.

According to TASC's Year-1 Implementation Timeline in its Strategic Plan document, "draft evaluation/assessment protocols" were to be approved by February 2003, "148 teachers from original 4 partner schools" would be "pre-/post-tested on content" as part of the February 6-7 and March 13-14 training workshops, and "students [would] take short test on unit content" in March 2003. (TASC (2002c), p. 4). In addition, by September 2003, NC DPI was expected to have developed "a trial set of test items related to 5^{th} and 8^{th} grade units" with the "items reviewed by TASC staff and partners but not administered". (TASC (2002c), p. 5). Thus, TASC also appeared to understand the necessity for an "assessment protocol" and possibly, by its reference to a "trial set of test items", the iterative nature of test development.

In TASC's Five-Year Implementation Plan, the following timetable (Table 15, below) was presented:

Table 15. From TASC's Five-Year Implementation Plan (emphasis added)

Time Period	Goals Addressed
10/1/02 – 6/30/03 (startup)	• about 3,552 students, K-5, of 148 teachers from partner schools will use a newly acquired curriculum unit for one goal of the NC Standard Course of Study (NCSCS). Science process skills and content knowledge tested pre/post. • . . .
7/1/03 – 6/30/04 (full academic year 1)	• about 25,200 students, K-8, of 540 teachers from partner schools (and additional partner school systems) will use a newly acquired curriculum unit for one goal of the NCSCS. Science process skills and content knowledge tested pre/post.
. . .	
7/1/04 – 6/30/05 (full academic year 2)	• about 50,400 students, K-8, of 1,080 teachers from partner schools will use a newly acquired curriculum unit for at least one goal of NCSCS. Some portion of these (up to half) may be using TASC-supplied curriculum units for two NCSCS goals (half the annual science curriculum). Science process skills and content knowledge tested pre/post each unit.
. . .	
7/1/05 – 6/30/06 (full academic year 3)	• about 75,600 students, K-8, of 1,620 teachers from partner schools will use a newly acquired curriculum unit for at least one goal of NCSCS. Some portion of these will be using TASC-supplied curriculum units for two or three NCSCS goals (up to 3/4 of the annual science curriculum). Science process skills and content knowledge tested pre/post each unit.
. . .	
7/1/06 – 6/30/07 (full academic year 4)	• about 100,800 students, K-8, of 1,620 teachers from partner schools will use a newly acquired curriculum unit for at least one goal of the NCSCS. Some portion of these will be using TASC-supplied curriculum units for from two to four NCSCS annual goals. Science process skills and content knowledge tested pre/post each unit. In this year, all 5th and 8th graders in the state will receive the NC statewide science test. Students in the TASC program will be compared with all other students.
. . .	
7/1/07 – 6/30/08 (full academic year 5) TASC free-standing w/o NSF MSP funds	• about 100,800 students, K-8, of 1,620 teachers from partner schools will use a newly acquired curriculum unit for at least one goal of the NCSCS. Some portion of these will be using TASC-supplied curriculum units for from two to four NCSCS annual goals. Science process skills and content knowledge tested pre/post each unit. All NC 5th and 8th graders again receive the statewide science test. Students in TASC compared with all other students.
. . .	

From this timetable, TASC appeared to understand that developing the tests, both the short curriculum unit tests and the statewide science tests, would take time.

Under the "Partner Roles and Responsibilities" section, TASC's Strategic Plan stated:

> . . . NCDPI has committed to the following:
> - development and implementation of evaluation and analysis protocols for impact on
>   - teachers' changes in content knowledge
>   - teachers' changes in attitudes toward inquiry-based teaching
>   - changes in student attitudes
>   - changes in student performance on end-of-grade in science, mathematics, and language arts
>   - student course choices
> - development of test items for statewide standardized testing that will measure student science process skills and content knowledge that may result from exposure to inquiry-based curriculum units, including those provided by TASC
> - development of short curriculum unit topic content tests comprised of a small set of knowledge level and problem-solving test items tied to the NC Standard Course of Study and the curriculum unit topics for grades K-8
>
> . . .
> (TASC (2002c), pp. 16-17; emphasis added).

Again, TASC appeared to understand that measurement professionals must be actively involved in the process of developing the tests.

Under the "Management Team Roles and Responsibilities" section, TASC's Strategic Plan stated:

> The management team consists of primary training staff, Program Manager, TASC Director, Associate Director for Training, Associate Director for Curriculum, Science Materials Center Manager, and temporary help for refurbishment during the three annual loan cycles. …
> (TASC (2002c), p. 18).

In this section, TASC enumerated the roles and responsibilities for each of the management team positions. Yet none of the responsibilities listed under each position included writing, *or reviewing*, items for the science tests.

The TASC year two scope of work for the NC DPI Instruction Evaluation and Assessment Team subcontract with TASC/Duke University continued to imply that TASC understood the necessity of measurement professionals' involvement in the creation of the tests, both the curriculum unit tests and the statewide science tests. This TASC document stated:

> With this agreement, the NC Department of Public Instruction contracts with Duke University's Teachers and Scientists Collaborating (TASC) to 1) assess TASC's impact on teachers and students …. To these ends, DPI will provide the following services:
>
> 1. develop test items suitable for statewide standardized testing, grades 5 and 8, that measure student science process skills and content knowledge potentially resulting from exposure to inquiry-based curriculum units, including those provided by TASC for 2004-05 (Investigating Weather Systems – grade 5, Motion & Design – grade 5, Landforms – grade 5, Solutions and Pollution – grade 8, Earth History – grade 8)
> 2. prepare and review short curriculum unit topic content tests comprised of a small set of knowledge level and problem-solving test items tied to the NC Standard Course of Study and TASC 2004-2005 curriculum unit topics for grades 3-8 (Soils – grade 3, Plant Growth and Development – grade 3, Human Body – grade 3, Magnetism and Electricity – grade 4, Food Chemistry – grade 4, Motion and Design – grade 5, Investigating Weather Systems – grade 5, Landforms – grade 5, Planetary Science – grade 6, Exploring Energy – grade 6, Thrill Ride – grade 7, Human Body Systems – grade 7, Solutions and Pollution – grade 8, Earth History – grade 8)
>
> . . .
>
> Payments beyond 50% of the Year 2 subaward amount will be contingent upon the TASC Director's approval. The TASC Director will serve as liaison, advise, and provide information on request. To begin measuring the above benchmarks, the NCDPI Evaluation & Assessment Team assumes responsibility for the following deliverables by the deadlines in the third column below [sic]:

Deliverable	Criteria
• assurances of validity and reliability of instruments and state (NC) test items • . . . • preliminary analysis of student and teacher data, . . . • preliminary overall testing design to measure changes in performance of students of participating teachers grades 5 and 8 (relevant to the projected 2007 test)	Data will include: • participating teachers' changes in knowledge of science content and process • . . . • . . . • . . . • . . . • Analysis will show differences among differences implementing TASC in different ways

(TASC 2004 subcontract with NC DPI; emphasis added).

These TASC documents (proposal, evaluation plan, implementation plan, 2004 subcontract with DPI) implied that TASC had an understanding of the test development process. That is, TASC appeared to understand that:

- the tests items needed to be aligned with instruction (i.e., NC Standard Course of Study and the curriculum units);
- measurement personnel (i.e., DPI) and content experts (i.e., TASC scientists) needed to work together to insure that the tests would be content valid;
- the tests needed to be reliable;
- the tests needed to be piloted before being used operationally;
- measurement personnel were needed to create and analyze data from the piloted tests;
- test development was iterative with results from pilot testing potentially impacting the operational tests; and

- developing the tests would take time.

In light of TASC's understanding of the test development process as documented above, one may have expected a more systematic development of the content framework in its subcontract with CERE/UNCG. That is, the content framework for the to-be-developed tests would have been substantiated by observations of training workshops (which observations were, in fact, included in the TASC evaluation plan), workshop (i.e., course) outlines, etc., such that each test blueprint was aligned with workshop curricular coverage. Instead, this researcher experienced a less than systematic development of content framework. Apart from information on TASC's website that stated each curriculum unit was aligned with specific NC SCS competency goals and objectives, no documentation was provided by TASC to substantiate each curriculum unit's alignment with specific NC SCS goals and object. Rather, at CERE's request, percentages of time spent on instructional objectives—necessary to construct the test blueprints—were assigned—based on subjective ("best guess") judgments of the TASC trainers.

In addition, and as reflected in its five year implementation plan, TASC appeared to understand that test development takes time. However, this understanding was not reflected in TASC's initial subcontract with CERE. In a February 21, 2005, email from the TASC project director, CERE was requested to perform the following tasks, by September 30, 2005:

- CERE to conduct item writing workshops for master teachers selected by TASC personnel; these teachers will write the items for student assessments and for teacher assessments

- CERE to create items for 23 modules:
    - modify/edit items
    - create assessments
    - pilot the assessments.
- CERE to perform item analyses on student assessments
- CERE to perform item analyses on teacher assessments

At a March 2005 meeting, the TASC project director gave priority to the Grades 3, 5, and 8 tests, reducing the actual number of tests from 23 to 10. Even so, the two-year delay caused by the previous test developer resulted in a compressed timeline—seven months (February through September 2005)—for CERE to develop ten operational tests.

Also unlike the TASC-NC DPI subcontract, the CERE-TASC subcontract never specified teacher or student "benchmarks" or that the tests were to be short (four or five items) content tests made up of knowledge-level and problem-solving items. As stated in the prior subsection, at the March 2005 meeting, the TASC project director indicated that each *operational* test was to include ten "medium to hard" (multiple-choice) items aligned with the NC SCS and that required examinees' higher order thinking skills.

Finally, rather than university science educators and teachers working together as item writers, TASC's subcontract with CERE included, as one of the deliverables, an item writing workshop that was to prepare TASC-trained teachers to write multiple choice questions that tapped examinees' higher order thinking skills. One major assumption implicit in this "deliverable" was that teachers would *want* to write items for these tests. Other assumptions implicit in this "deliverable" were that (1) teachers would know their

science content well enough to write the items, and (2) teachers would be able to write multiple-choice questions measuring higher-order skills—all within a very short timeframe. As will be elaborated upon in the next section (phase 3), none of these assumptions was correct.

In summary, the factors that affected phase 2 (test specifications) of the test development process included:

1. National and state science standards: These standards emphasized the active nature of science instruction—i.e., inquiry-based—which, in turn, was central to TASC's teacher training. This inquiry-based emphasis affected the tests' blueprints in that the tables of specifications were to reflect both NC SCS instructional objectives *and* NC thinking skills; and test items were to be multiple-choice, were to require examinees' higher level thinking skills (i.e., no knowledge-level items), and were to be of "medium to hard" difficulty.
2. TASC's understanding of the role of the to-be-developed tests in its project evaluation: TASC's initial understanding—articulated in its project evaluation, in its implementation plan document, and in its subcontract with NC DPI—was that TASC expected to use test results from short curriculum unit content tests and from large assessment state science tests as quantitative evidence to its funding agency (NSF) that TASC was meeting its stated teacher and student outcome goals.

   This initial understanding of the role of the to-be-developed tests appeared to change somewhat by 2005 when TASC subcontracted CERE to develop separate tests—ten for students and ten for teachers—on science content and science process

tied to the NSF-approved curriculum units that were tied to the NC SCS. While the purpose of the to-be-develop tests had not changed from TASC's original intentions as stated in its evaluation plan—that is, to measure improvements in content knowledge of participating teachers and their students, the use of the test results apparently *had* changed from providing quantitative evidence that TASC was meetings its teacher and student outcome goals to providing "evidence from [workshop] training".

3. TASC's understanding of the test development process: TASC documents (proposal, evaluation plan, implementation plan, 2004 subcontract with DPI) articulated what appeared to be TASC's understanding of test development. That is, TASC appeared to understand that:
    - the tests items needed to be aligned with instruction (i.e., NC Standard Course of Study and the curriculum units);
    - measurement personnel (i.e., DPI) and content experts (i.e., TASC scientists) needed to work together to insure that the tests would be content valid;
    - the tests needed to be reliable;
    - the tests needed to be piloted before being used operationally;
    - measurement personnel were needed to create and analyze data from the piloted tests;
    - test development was iterative with results from pilot testing potentially impacting the operational tests; and
    - developing tests takes time.

Based on TASC's apparent understanding of the test development process, the expectations were that the development of the content framework would have been very systematic (it was not), that sufficient time would have been allotted for the development of ten operational tests (it was not, due in large part to the two-year delay caused by the previous test developer), and that the implicit assumptions about item writers were valid (they were not).

Pilot Test (Phase 3)

This phase includes the construction and evaluation of the pilot test(s) and includes such tasks as:

- generating items;
- selecting items based on:
  - content quality and scope,
  - instructional objective addressed,
  - cognitive skill to be used by examinee, and
  - appropriateness of the item for population of intended test-takers;
- assembling items into pilot test;
- administering pilot test to subset of intended population of test-takers;
- evaluating/revising items from piloting test (i.e., item analysis);
- evaluate scoring procedures; and
- evaluating/revising test administration procedures.

The following sections—item development, test assembly, test administration, data analyses, and test revision—describe the sequence of events that occurred in this

particular phase of test development. Each section includes a discussion of the factors that affected it.

Item Development

As stated in the phase 2 discussion, TASC initiated CERE's item development tasks prior to completion of its (TASC's) test specifications tasks. TASC expected that TASC-trained teachers would write items for the tests. Expecting expedited approval of the TASC-CERE subaward, this researcher began preparation in January 2005 for the item writing workshops that, according to the *proposed* scope of work, were to be held in mid-February 2005. The expected pattern for item generation is presented in Table 16. It is worth noting that the resources for this task included a *signed* subcontract between TASC and CERE as well as TASC-"nominated" (versus pre-selected) teacher-item writers.

Table 16. Expected pattern for item generation task

Resources	Activities	Outputs	Outcomes
• Testing standards • Signed TASC-CERE subcontract • TASC-nominated teacher-item writers • TASC scientists (as teacher-trainers and as item reviewers) • Curriculum units • NC Standard Course of Study • NC Thinking Skills • Tables of specifications	• MC Item Writing Workshop → MC Item Writing Workbook	• Higher order (i.e., above knowledge level) MC Qs: o 8 usable MCQs for teacher test o 8 usable MCQs for student test	• Teacher tests of 10-15 MCQs • Student tests of 10-15 MCQs

On February 21, 2005, TASC emailed CERE with a revised description of work and budget justification for the proposed subaward between TASC and CERE. The period of performance for the proposed subaward was February 15, 2005 through September 30, 2005—a 7.5 months period of performance. With no signed subaward by February 2005, this project was behind schedule from the start.

Recruitment of Teacher-Item Writers—Part 1

On March 21, 2005, the TASC project director emailed the project evaluator and the CERE acting director a "list of [48] teachers, by kit, from Iredell and Alamance, with email addresses that our [TASC] trainers nominated to attend the item writing training. We also included a few from Chatham and Orange who are especially good." The TASC project director instructed CERE to "go ahead and contact these folks by email."

TASC criteria for selecting potential item-writers included: (1) teachers who had received prior TASC training on at least one science curriculum workshop, (2) TASC-trained teachers who had used the science unit in their classrooms, *and* (3) TASC-trained teachers who attended one three-hour item writing workshop, prepared and presented by this researcher, at the TASC Training Center in Durham, NC. In addition, TASC stipulated two item writers per science unit.

On March 22, 2005, the TASC project director, TASC project evaluator, the CERE acting director, and this researcher met to discuss the (unsigned) subaward. At this meeting, the TASC project director, in his response to questions asked by this researcher, gave priority to the following nine curriculum units for which tests were to be developed:

Grade 3
- Human Body (NCSCOS Goal 4, subgoals 1 through 5)
- Plant Growth & Development (NCSCOS Goal 1, subgoals 1 through 6)
- Soils (NCSCOS Goal 2, subgoals 1 through 6)
- Investigating Objects in the Sky (NCSCOS Goal 3, subgoals 1 thtrough 6)

Grade 5
- Investigating Weather Systems (NCSCOS Goal 3, subgoals 1 through 6)
- Landforms (NCSCOS Goal 2, subgoals 1 through 7)
- Motion & Design (NCSCOS Goal 4, subgoals 1 through 7)

Grade 8
- Earth History (NCSCOS Goal 1, subgoals 1 through 6, 8; Goal 2, subgoals 1, 3; Goal 5, subgoals 1 through 5)
- Solutions & Pollution (NCSCOS Goal 1, subgoals 1 through 5, 8; Goal 3, subgoals 1, 7, 8; Goal 4, subgoals 4, 5)

In addition, the TASC project director indicated that each operational test was to be comprised of ten multiple-choice items that were to be of "medium to hard" difficulty and that required examinees' higher-level thinking skills (i.e., no knowledge-level items).

To accomplish the work under this (unsigned) subaward between TASC and CERE, this researcher documented a tentative test development timetable as follows:

March 2005 — Recruit item writers:
- 2 item writers/curriculum unit—no overlap (exception: Plant Growth & Development, one item writer)
- item writer to have received training on curriculum kit on which being asked to write items

April 13, 2005 — Conduct Item Writing Workshop
- *preferably* ONE workshop for ALL participating teachers
- participants be asked to submit a *minimum* of five questions for teachers and five questions for students
- to be held at UNCG (MacDonald lounge ??)
- date: to be determined

	• length of workshop: approximately 3 hours • at close of workshop, participants to be reminded, per their letter, to submit multiple choice test questions NLT 10 workdays after end of workshop
April 23, 2005	Receive test questions from item-writers
April 23-30, 2005	Review/revise/format questions for pilot testing
May 2-6, 2005	Pilot testing: (only time to do this since EOG testing occurs 5/10-12/05 and EOC testing occurs 5/24--27/05): • on TASC-participating classrooms • on non-participating classrooms matched to participating classrooms • dates for pilot test: first week of May is only time can do this during 2004/05 academic year
May 8-31, 2005	Analyze results from pilot testing
June-July 2005	Revise test questions as needed
August 2005	Assemble tests for re-piloting in 2005/2006 academic year
September 2005	• Beginning of month, administer revised pilot tests, as needed • Analyze results from revised pilot tests
October 2005	Revised questions from revised pilot tests, as needed
November 2005	Prepare tests for operational administration
December 2005/January 2006	Administer operational science tests to teachers and to students in TASC-participating and non-participating classrooms
January 2006	• Analyze test data. • Write up results in final report to TASC.

Of the 48 teachers on TASC's list, only nine had the requisite training on two units and of those nine, only six had training on one of the third, fifth, and/or eighth grade

units. The TASC project director selected teachers from the list of 48 to bring the total number of TASC recommended "master" teachers up to 15: two teachers for Grade 3 Human Body, one teacher for Grade 3 Plant Growth & Development, two teachers for Grade 3 Soils, two teachers for Grade 5 Investigating Weather Systems, two teachers for Grade 5 Landforms, two teachers for Grade 5 Motion & Design, two teachers for Grade 8 Earth History, and two teachers for Grade 8 Solutions & Pollution.

Upon returning to CERE from the March 22, 2005 meeting, this researcher sent each of the 15 TASC-recommended teachers an email that:

- explained the role of CERE in helping to develop science tests as part of TASC's evaluation;
- acknowledged the TASC training they had received in the use of the curriculum kit materials;
- indicated they had been recommended by TASC personnel as an item writer candidate; and
- requested their participation as an item writer for the specific science unit upon which they had been trained.

In addition, the email explained their responsibilities, if they chose to participate:

1. Attend an item writing workshop that was expected to be held on Wednesday, April 13, from 4-7 pm.
2. Submit, within ten working days of the item writing workshop, a minimum of ten multiple choice questions, five of which would be used for student tests and five of which would be used for teacher tests.

The email explained that upon the submission, review, and revision (if applicable) of the ten items, the participant would receive a check for $50.

Subaward status: On March 23, 2005, the TASC program director emailed CERE that Duke's program officer had approved the CERE/TASC subaward. However, this did not mean that the Duke Office of Contract and Grants had signed the subaward. The project was now six weeks behind schedule.

On March 25, 2005, this researcher notified the TASC program director by email that only 3 of the 15 teachers had responded to the March 22 email, indicating they would be unable to attend the April 13 workshop. By March 29, 2005, only one teacher expressed an interest in participating as an item writer but only if she could carpool with others in her school district.

Between March 25 and April 6, 2005, this researcher and the TASC project director discussed a back-up plan for the item writing workshop should the April 13 workshop not take place. The TASC project director said that in June 2005, TASC would conduct Trainers' Institutes at the TASC Training Center in Durham, NC. At these institutes, TASC-trained teachers would learn how to train other teachers on the use of the science curriculum kits. This researcher suggested adding the item writing workshop at the end of the institutes and offered to present the item writing workshop at the TASC Training Center because the teachers were at that facility making it convenient for them to attend the workshop. This meant adding an extra half-day to the teachers' training as the item writing workshop would be approximately 3 1/2 hours long.

On April 4, 2005, this researcher sent a follow-up email to the 10 (of the 15) teachers who had not responded to the March 22 email. By April 6, 2005, of the 15 teachers emailed on March 22, one responded "yes", one responded "maybe", nine responded"no", and four never responded. This researcher relayed this information to the TASC project director who responded "I doubt things will change much by the end of the today. So, if we were to add the item writing workshop to the trainer's institutes in June, how do you think that should be done? … The item writing workshop would need to be held on a continguous day. …"

In an April 8, 2005, email to the TASC program director, this researcher indicated that the April 13 item writing workshop would not be held due to lack of positive responses. In addition, this researcher initiated the next attempt to recruit teachers as item writers.

Recruitment of Teacher-Item Writers—Part 2

In the April 8, 2005 email to the TASC project director, this researcher offered to present the item writing workshop four times—the day after each of the four two-day institute trainings—June 16, 17, 23, 24 from 9 am to 12:30 pm. She pointed out that trainees would receive item writing training *after* they had received curriculum unit training, that they would not have to travel elsewhere to receive the training, and that if the item writing workshop were held four times, eight of the nine curriculum units given priority by TASC would be covered (exception: "Investigating Objects in the Sky"). Lastly, this researcher noted that the participants would receive 2.0 "renewal credits" for attending the TASC training and asked whether it would feasible to offer 0.5 "renewal

credit" for those who attended the item writing workshop and submitted a total of 16 *usuable* questions (8 for teachers and 8 for students), where "usable" was defined as questions that had been reviewed, revised, and accepted by CERE. In his response, the TASC project director indicated his approval in offering the item writing workshop four times on June 16, 17, 12, and 24 and in offering 0.5 renewal credits for the three-hour workshop and 16 questions accepted by CERE.

Subaward status update: On April 11, 2005—almost two months after the proposed start date of February 15, 2005, UNCG's Director, Office of Sponsored Programs signed the subaward between CERE/UNCG and TASC/Duke University. The subaward's cover sheet indicated a January 1, 2005 start date and a September 30, 2005 end date. However, by early May 2005, Duke's Office of Contract and Grants had not yet signed the subaward. For this researcher, who had been working on the TASC-CERE subaward since January 2005, to be paid by UNCG, CERE had to file an "Assumption of Risk".

On May 4, 2005, this researcher emailed the TASC project director asking about the status of a file from the TASC training director with the list of registrants' names and mailing addresses for the June institutes. This researcher indicated that she wanted to send the CERE letter about the item writing workshops separate from the TASC mailing about the training institutes. The TASC project director responded by indicating that TASC would send the list with the registrants' names and mailing addresses to CERE on Monday, May 9, or Tuesday, May 10. He indicated that TASC had extended their

enrollment deadline to May 2 because enrollment had been light and late with 85 out of 120 openings filled. In addition, he noted:

> 1) only 7 middle school teachers have signed up and a middle school sessions needs 10 …. If we don't get 3 more, we won't hold either of the middle school sessions. Those were to be on June 20-22 and June 21-23. If we get 10 signed up, we will hold only 1 of those two dates, probably the one on the 20th-22nd.
> 2) about a half dozen application forms were faxed to us by central office staff in various districts. Few of those faxed forms include teachers' home addresses. The best you can do on these is send your letter to the schools. As you know, that's unreliable.

On May 11, 2005, this researcher emailed the TASC project director asking 1) whether TASC had closed the registration for the June training sessions, 2) whether some sessions had been dropped due to low enrollment, and 3) when to expect the file with each workshop's registrants' names and addresses. The TASC project director responded by emailing the contact list for the two workshops TASC would hold in June 2005. Keeping in mind that Duke's Office of Contract and Grants had yet to sign the TASC-CERE subaward, the TASC project director also stated:

> Today, we submitted a request to add … to your budget under the consultant services line [to be able to pay teachers a stipend]. … Our office of research support will write a new agreement, which Duke will sign and send to your contracts office for signature. I tell you this now so that you can add whatever incentives are reasonable to get teachers to write the items we need.
>
> Regarding placement of the added funds in the consultant services line, I am told that you will be free to use them to pay teachers for their work in writing items, but please let me know if this is not the case and you need some or all of it in participant support. We thought putting it in that line gives you more flexibility in the event that teachers do not respond and we need to go to a 'plan B' or 'C'. (Personal communication; emphasis added.)

In a May 16, 2005, email from the TASC project director, he stated:

> In the additional budget coming to you, I've requested that all of the new funds go under the participant support line. There is no problem getting you the money, so include any stipend you need to as enticement in your letter. I'll have word on the budget line issue as soon as the administrator at Duke calls me back, should be later today.

On May 16, 2005, this researcher mailed letters to the 25 registrants for the June 13-15 training institute and to the 16 registrants for the June 14-16 training institute. The letter explained the role of CERE in helping to develop science tests as part of TASC's evaluation, acknowledged the TASC training they had received in the use of the curriculum kit materials, and requested their participation as an item writer for the specific science unit upon which they had been trained. In addition, the letter explained their responsibilities, if they chose to participate: attending a three-hour item writing workshop and submitting a minimum of 16 (initial) multiple choice test questions, half of which would be written for student tests and half for teacher teachers, within three weeks of the item writing workshop. Last of all, the letter explained that upon the submission, review, and revision (if applicable) of the 16 initial items, the participant would receive "at least $200 and one renewal credit." A stamped, self-addressed postcard was provided for each recipient to indicate their desire to participate or not participate. The letter requested the postcard be mailed back to CERE by Wednesday, May 25th. On May 18, 2005, this researcher mailed an identical letter to 22 teachers who had previously received TASC institute training.

On May 24, 2005, this researcher emailed the TASC project director to let him know that as of May 23, 2005, four June 14-16 registrants and two June 13-15 registrants had agreed to participate as item-writers.

Following the chronological sequence of events, the discussion moves back briefly to phase 2 (test specifications) of test development. On May 25, 2005, this researcher emailed the TASC project director requesting that he and his team "provide the percentage of time devoted by the curriculum unit to *each* NCSCS instructional objective by Competency Goal by Grade." To facilitate TASC's response, this researcher included tables for each NCSCS (Science) Competency Goal for Grade 3 (goals 1-4), Grade 4 (goals 1-4), Grade 5 (goals 1-4) and requested that they be returned before June 3, 2005. On May 31, 2005, the completed tables were returned by email to this researcher, who then used them to create tables of specifications to be included in the item writing workshop materials.

On June 2, 2005, in response to the TASC project director's telephonic request, this researcher prepared a brief status report with test development information for TASC to include in their June 30 annual report to NSF. Part of the report stated:

> Since one of the outcomes of NC education is that students learn to think critically and creatively, and since the NC Standard Course of Study provides the competencies that students should demonstrate in science, all test questions will be classified by two dimensions:
> (1) by the specific Instructional Objective being measured by the question, and
> (2) by the NC thinking skill(s) (adopted by the NC Department of Public Instruction as their model to classify questions for NC tests) the student will utilize to correctly answer the question.

> In order to develop test items for these science assessments, the Center has recruited elementary school teachers, who have been trained by TASC on at least one curriculum unit, as item writers. In June 2005, the teacher-item writers will attend an item writing workshop prepared and presented by the Center at which the teachers will be instructed in the creation of effective multiple choice test questions. After attending the item writing workshop, the teacher-item writers will write multiple choice test questions for their particular curriculum unit.
>
> These test questions will be submitted by July 2005 to the Center where the items will reviewed for format, clarity, etc. Once reviewed, and revised if necessary, the items will be assembled into elementary grade level science tests which TASC will pilot test some time in August/September 2005 in selected TASC schools. Data from the pilot test will be submitted by TASC to the Center for appropriate analyses. Based upon results from these analyses, the Center will make recommendations to TASC regarding test revisions. Once revised, TASC then will be able to administer these science tests operationally to all TASC schools. Through the use of these science tests, which will be administered in TASC classrooms prior to the use of the NSF-approved curriculum units and upon the completion of the curriculum units, TASC anticipates finding improvement in teachers and students science content and process knowledge.

On June 6, 2005, after reviewing the report, the TASC program director requested that it be updated, after this researcher met with the item writers, to reflect the number of item writers, what they would be working on, and the deadline for items.

Subaward status update: On June 8, 2005, after working on the TASC grant since January 2005, this researcher received her appointment letter as project manager of the TASC test development grant "for the period of May 1, 2005 through September 30, 2005, at a salary . . . based on five months of one-fourth-time service . . . ." On June 11, 2005, six months after work began, the subagreement (#05-SC-NSF-1057) between CERE/UNCG and ASC/Duke University was signed by Duke's Office of Research Support. On the same day, Modification 01—adding enough money for item-writer stipends to incentivize teachers to participate—to this subagreement was signed by

Duke's Office of Research Support. This modification was signed, on June 15, 2005, by UNC-G's Office of Sponsored Programs.

Item Writing Workshop(s)

On June 16 and 17, 2005, this researcher conducted two half-day item writing workshops at the TASC Training Center. Each workshop covered the following:

- Teachers and Scientists Collaborating Project
- Center for Educational Research & Evaluation
- Test Development Process
- Test Purposes
- Essential Test Characteristics
- Table of Specifications
- Models of Thinking Skills
- Item Types
- Item Construction
- Multiple Choice Items
- Item Writing Guidelines
- Item Writing Workbook

Materials that this researcher prepared and provided to each teacher included:

- a copy of the Powerpoint presentation;
- a *Multiple Choice Item Writing Workbook* with Appendix A (prepared by this researcher; see Appendix E); and

- "Directions to Item Writers" with the following attachments:
    - a Student Science Item Specification Sheet
    - a Teacher Science Item Specification Sheet
    - test blueprints that included the number of items to be written for each NCSCS instructional objective to be covered by the test.

Table 17 indicates the number of attendees (with total registrants listed in parentheses) for the two workshops.

Table 17. Number of attendees at June 2005 item writing workshops

	June 16 Workshop Attendees	**June 17 Workshop Attendees**
Gr 3 Human Body	2 (2)	
Gr 3 Soils	2 (3)	
Gr 3 Plant Growth & Development	0 (1)	6 (6)
Gr 4 Food Chemistry		1 (2)
Gr 4 Magnetism & Electricity	5 (5)	
Gr 5 Investigating Weather Systems	0 (1)	
Gr 5 Landforms		2 (2)
Gr 5 Motion & Design		2 (3)
Totals	9 (12)	11 (13)

At the end of each workshop, this researcher asked attendees to anonymously respond to the following questions:

- What did you find most helpful?
- What did you find least helpful?
- Any suggestions as to how to improve it?

Of the 18 (out of 20) attendees who responded, 13 found information included in the workbook to be most helpful, 8 found parts of the Powerpoint presentation to be least helpful, and 10 suggested more "time to work together and practice writing a few questions".

On June 27, 2005, this researcher emailed an updated status report to the TASC program director. The last two paragraphs from the June 2 report were revised as follows:

> In order to develop test items for these science assessments, the Center has recruited elementary school teachers, who have been trained by TASC on at least one curriculum unit, as item writers. In June 2005, the Center conducted an item writing workshop at the TASC Training Center. Twenty teacher-item writers attended and were presented with requisite information for the creation of effective multiple choice test questions. These teacher-item writers are writing multiple choice test questions for their particular curriculum unit.
>
> These test questions will be submitted by July 2005 to the Center where the items will be reviewed for format, clarity, etc. Once reviewed, and revised if necessary, the items will be assembled into elementary grade level science tests which TASC will pilot test some time in August/September 2005 in selected TASC schools. Data from the pilot test will be submitted by TASC to the Center for appropriate analyses. Based upon results from these analyses, the Center will make recommendations to TASC regarding test revisions. Once revised, TASC will administer these science tests operationally to all TASC schools. Through the use of these science tests, which will be administered in TASC classrooms prior to the use of the NSF-approved curriculum units and upon the completion of the curriculum units, TASC anticipates finding improvement in teachers and students science content and process knowledge.

After the workshops and prior to June 27, 2005, one of the Grade 4 Magnetism & Electricity item writers brought to the TASC program director's attention that the ten percent of time spent on objective 3.2 was incorrect. In the TASC project director's

response to the item writer, which he also sent to this researcher, he acknowledged that no time was spent on this instructional objective (and thus no items needed to be written for it).

On June 28, 2005, this researcher sent an email to all item-writers in which she provided a correction to the test blueprint for the Grade 4 Magnetism and Electricity competency goal. In addition, this researcher indicated that practice items submitted by a few item writers had been written at the knowledge-level of the NC Thinking Skills taxonomy. She stated that while information on the knowledge-level was included in the item writing workbook (to provide item writers with complete information on the NC Thinking Skills taxonomy), it was stated at the workshops that the test blueprints were the "guides for item writing" and that "none of the blueprints" included the knowledge-level thinking skill because "we are interested in determining whether students/teachers know how to *use* science content/process skills."

Factors that affected teacher-item writer recruitment.

TASC's operational understanding of the test development process appeared to be much weaker than its conceptual understanding that was documented in the phase 2 discussion. Briefly, from its proposal, evaluation plan, implementation plan, and 2004 subcontract with DPI, TASC appeared to understand that:

- the tests items needed to be aligned with instruction (i.e., NC Standard Course of Study and the curriculum units);
- measurement personnel (i.e., DPI) and content experts (i.e., TASC scientists) needed to work together to insure that the tests would be content valid;

- the tests needed to be reliable;
- the tests needed to be piloted before being used operationally;
- measurement personnel were needed to create and analyze data from the piloted tests;
- test development was iterative with results from pilot testing potentially impacting the operational tests; and
- developing tests takes time.

TASC's weak operational understanding of the test development process affected the teacher-item writer recruitment efforts. From the very beginning of this project, TASC grossly underestimated the time required to recruit teacher item-writers. The CERE-TASC subcontract, that was not fully authorized until Duke signed it in June 2005, had no time allotted for teacher recruitment. This would have been understandable had TASC pre-selected their teacher-item writers prior to February 2005. However, TASC did not pre-select teacher item-writers.

In addition to its gross underestimation of recruitment time, TASC assumed that its participating teachers would be understanding and responsive to TASC's need for pre- and post-test data that could be matched by student and teacher. In an August 24, 2006 interview, this researcher asked the TASC project director his thoughts concerning why the initial recruitment attempt of teachers as item writers failed. He stated:

> … I don't think the teachers are committed to TASC to the program. I think the teachers like what they're getting—a lot—but it's a one-way street. They're not really thinking about giving back because … they're just full out occupied by trying to do their jobs—especially the good ones. The more skilled the teacher is

> the more they're leaned on in the school districts. So … this was … really an outside extra thing ….

TASC also did not appear to take into consideration how threatening pre- and post-testing may seem to classroom teachers, irrespective of telling the teachers that the test results were to be used to evaluate the *project*, not the teacher. Instead, TASC assumed that its participating teachers would *want* to attend a three-hour item writing workshop and write questions for tests that would be administered to them and their students, with little remuneration (initially $50) for their efforts.

Item Generation and Revisions

After the item writing workshops were conducted, teacher item-writers were given three weeks (until July 8, 2005) to submit their initial 16 items—8 teacher test items and 8 student test items—each with an item specification sheet. All items were to be written according to the test blueprint provided to each item writer at the workshop.

From July 6 through July 13, 2005, this researcher received initial items from 13 of the 20 item writers who attended the item writing workshops. Even this step was not without problems. For instance, on July 11, item writer (IW) 15 simply emailed to this researcher a list of 25 short multiple choice questions. A sample of these questions included:

1. The key element that interacts with a magnet is
   A. gold.
   B. iron.
   C. lead.

3. To repel is to
    A. come together.
    B. push away.
    C. remain stationary.

6. A source of electric energy is a
    A. d-cell.
    B. switch.
    C. circuit.

Because none of the items had a requisite item specification sheet, there was no way to know whether an item was for the student or teacher test, what instructional objective was covered by the question, or what the correct answer was to the question. This researcher emailed IW15 on July 12, acknowledging receipt of the 25 questions and requesting a completed item specification sheet for each question. Item specification sheets were never received from IW15, even after subsequent emails from this researcher.

IW10 and IW03 emailed this researcher, on July 11 and July 14, respectively, that they were still working on their questions. Also on July 11, IW16 emailed this researcher that she had dropped out, stating:

> as I began to work on the questions I realized how difficult it was to use only one content goal at once, not have it knowledge based, and have the answers match those goals provided. After spending 3 hours on one question without success in completion, I decided this is not going to work with me finishing 16. I have attempted about 3 different questions that could be motified [sic[for teachers making it a possible 6 out of 16, but none of them are at a point where [sic] I would feel comfortable handing them in. sorry I couldn't help.

On July 14, 2005, this researcher sent a follow-up email to the remaining four workshop attendees who had not submitted any initial items. Two responded that they were still working on their items; two never responded.

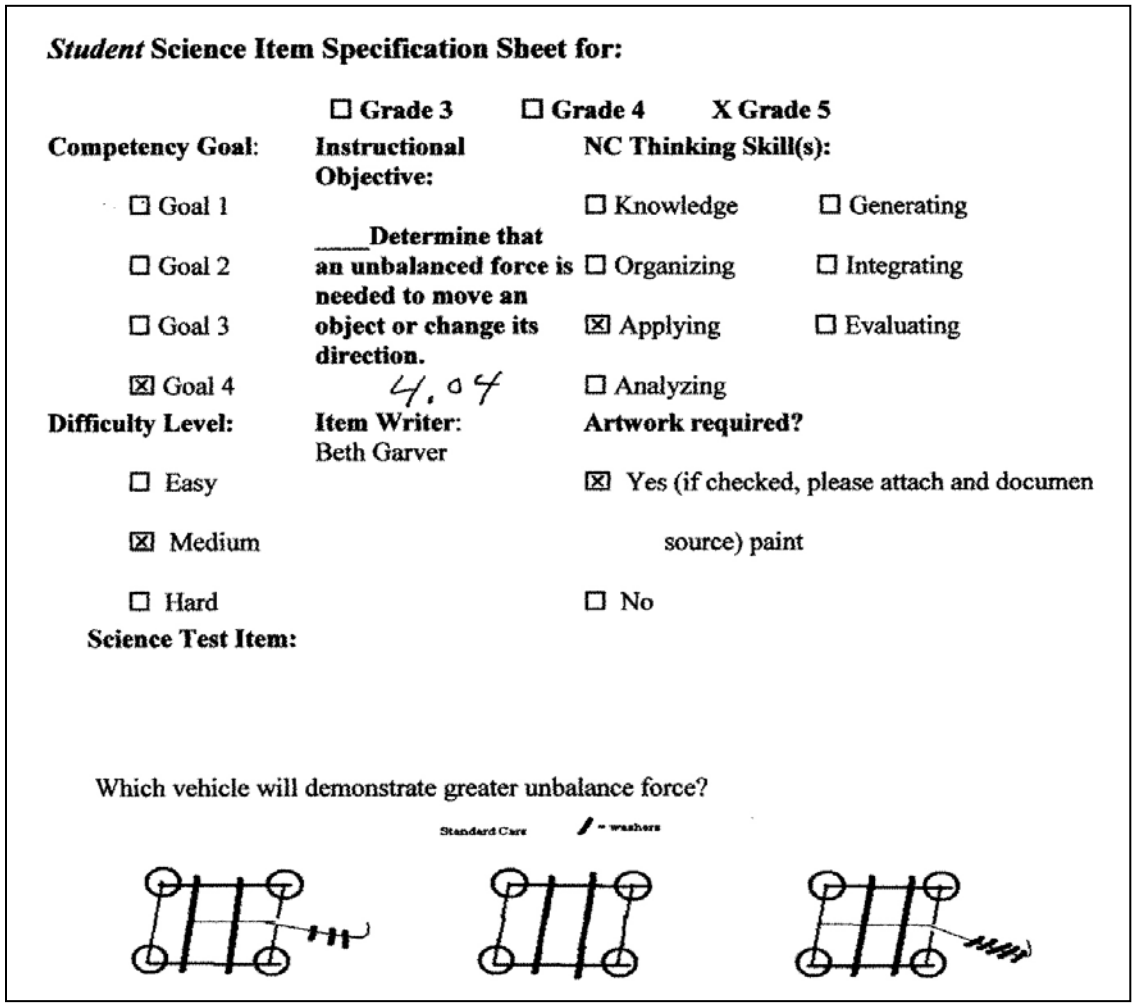
Throughout July 2005, this researcher received and reviewed initial items, verifying that items were written according to the item writer's test blueprint. Item quality was evaluated using criteria from Table 2 in the *Multiple Choice Item Writing Workbook* (multiple choice item writing checklist; reproduced below) that was provided to each workshop attendee.

Table 2 (from *MC Item Writing Workbook*). Multiple choice item writing checklist

1. General	Yes	No
a. Is the wording of the item clear and unambiguous?		
b. Does the item present one--and only one--problem?		
c. Is the item written at appropriate reading level for all students?		
d. Does each item measure only one instructional objective?		
e. Have punctuation, capitalization, spelling, and grammatical structure of the item been checked?		
f. Does the item avoid culture-specific references?		
2. The Stem	Yes	No
a. Is the problem stated concisely as a complete statement/ question?		
b. Is the stem presented positively?		
c. Are the directions in the stem clearly stated?		
d. Have extraneous cues to the correct answer been avoided?		
3. The Alternatives	Yes	No
a. Is there one--and only one--clearly correct answer?		
b. Is the correct answer supported by documentation (and not an expression of opinion)?		
c. Are the incorrect alternatives logical and plausible and unlikely to be eliminated by someone who does not know the material?		
d. Are the alternatives grammatically consistent with the stem?		
e. Is the correct response about the same length as one or more of the distractors and not any more technical than the other responses?		
f. Have *none-of-the-above, all-of-the-above,* or *I-don't-know* been avoided as alternatives?		

After finding a few items that did not correctly state the facts (e.g., a hand-drawn picture of a parallel circuit that was not a parallel circuit), this researcher obtained, from UNCG's Teaching and Learning Center and from a variety of science education websites, science content materials to attempt to verify the accuracy of item content. In addition, this researcher requested that items be revised when what was being asked of an examinee was unclear. For example, one student test question initially submitted by IW20 is included in Figure 7, below. This researcher, in her July 16 email to IW20, requested revision of this item, stating: "The stem is unclear, i.e., there is no context; there is no indication of what 'washers' represent; there is no way of determining 'force,' unbalanced or otherwise."

This process of verifying the accuracy of item content, reviewing initial questions, verifying that items were written according to the test blueprints, requesting revisions, and reviewing revised and re-revised questions continued throughout July 2005. Only one item writer submitted his/her final items by July 20.



***Student* Science Item Specification Sheet for:**

☐ Grade 3 ☐ Grade 4 X Grade 5

Competency Goal:

☐ Goal 1
☐ Goal 2
☐ Goal 3
☒ Goal 4

Instructional Objective:

____Determine that an unbalanced force is needed to move an object or change its direction.

4.04

Item Writer: Beth Garver

NC Thinking Skill(s):

☐ Knowledge ☐ Generating
☐ Organizing ☐ Integrating
☒ Applying ☐ Evaluating
☐ Analyzing

Difficulty Level:

☐ Easy
☒ Medium
☐ Hard

Artwork required?

☒ Yes (if checked, please attach and documen source) paint
☐ No

Science Test Item:

Which vehicle will demonstrate greater unbalance force?

Standard Cars = washers

Figure 7. Initial multiple choice question by item writer 20.

On August 2, 2005, this researcher emailed the TASC project director that out of 20 item writers, 5 had dropped out, 2 or 3 item writers submitted items that required little revising, and items submitted by the remaining item writers typically required much work. This researcher provided the TASC project director with the following "status report":

Here is where we stand as of today regarding the TASC tests:

3rd Grade Units:

- Human Body
  - IW01: dropped out (since never responded to anything, including 7/14 follow-up email)
  - IW02: 7/28 emailed her revised items to me; I'm in the process of reviewing them …
- Plant Growth & Development
  - IW03: as of 7/14 email, still working on items; have received nothing to date
  - IW04: waiting for revised items; received 8/2 via fax; in process of reviewing them
  - IW05: waiting for revised items [dropped out, 8/2]
  - IW06: 8/1 dropped out
  - IW07: as of 7/14 email, still working on items; have received nothing to date
  - IW08: 7/28 dropped out
- Soils
  - IW09: dropped out (since never responded to anything, including 7/14 follow-up email)
  - IW10: as of 7/11 email, still working on items; have received nothing to date [emailed me later—to receive items by 8/19]

4th Grade Units:

- Food Chemistry
  - IW11: 7/29 fax'd her revised items to me; I'm in the process of reviewing them (2 minor revisions needed)
- Magnetism & Electricity
  - IW12: 7/14 emailed me asking for more time; no items received to date
  - IW13: 7/20 received "usable" items
  - IW14: 8/2 received revised items; I'm in the process of reviewing them
  - IW15: 7/11 emailed 25 questions with *no* item specification sheets attached; I emailed her 7/12 requesting item specification sheets for each item; I have received no response to date
  - IW16: dropped out 7/11

5th Grade Units:

- Landforms
  - IW17: working on revised items
  - IW18: working on (one) student item needing revision; will email teacher items to be today or tomorrow [8/3 student Qs re c'd; in process of reviewing them; waiting on tchr items]

- Motion & Design
  - IW19: 8/1 emailed revised items; I'm in the process of reviewing them[--4 minor revisions needed]
  - IW20: working on revised items [8/3 rec'd; in process of reviewing them]

The TASC project director responded that he was "disappointed to see this" and that if the item writers did not submit their items before school, "we might not get them at all". He also indicated that one of the item writers had called him to say "she was meeting a lot of 'roadblocks' in the writing process, and that writing them was turning out to be a lot harder than she thought. She also said she knew of others who are having trouble. She said she didn't know if they need more training, or what."

In responding to the TASC project director's email, this researcher indicated that the "roadblocks" mentioned by the item writer may have been that "most of the teachers did not expect that writing *good* multiple choice items—at greater than knowledge level—was difficult." She expressed uncertainty that additional training would help because, as she stated at the workshop, writing multiple choice items was difficult because one had to "juggle" multiple priorities, such as content, higher order thinking skills, formatting, grammar, etc. In closing, this researcher indicated she "would be happy to sit down with you to show you some of the initial items I received and my responses to the item writers".

The iterative process of review-revise-review revisions-revise revisions continued with the item writers through August 2005. On August 19, 2005, this researcher emailed the TASC project director with an update as to the status of the item-writing. The results,

summarized in Table 18, indicated that four item writers had completed their task and that ten item writers continued to work on their test questions.

Table 18. Status of item writing as of August 19, 2005

Item Writer	Items written for:	Status of "usable" items
IW02	Gr 3: Human Body	Received 8/8/05
IW04	Gr 3: Plant Growth & Development	Received 8/3/05
IW13	Gr 3: Magnetism & Electricity	Received 7/20/05
IW14	Gr 4: Magnetism & Electricity	Received 8/8/05/05
IW03	Gr 3: Plant G &D	Still working on
IW07	Gr 3: Plant G & D (a potential drop out since I have not yet received any items from her even after following up via email)	Still working on
IW10	Gr 3: Soils	Still working on
IW11	Gr 4: Good Chemistry	Still working on
IW12	Gr 4: Magnetism & Electricity	Still working on
IW17	Gr 5: Landforms	Still working on
IW18	Gr 5: Landforms	Still working on
IW19	Gr 5: Motion & Design	Still working on
IW20	Gr 5: Motion and Design	Still working on
IW15	Gr 4: Magnetism & Electricity	Still working on
IW05	Gr 3: Plant Growth & Development	Dropped out 8/2/05
IW06	Gr 3: Plant Growth & Development	Dropped out 8/1/05
IW08	Gr 3: Plant Growth & Development	Dropped out 7/28/05
IW01	Gr 3: Human Body	Dropped out/no responses
IW09	Gr 3: Soils	Dropped out/no responses
IW12	Gr 4: Magnetism & Electricity	Dropped out 8/19

In addition, this researcher requested direction from the TASC project director as to whether a "final" due date should be set for all usable items to be submitted. In his August 22, 2005, response, the TASC project director stated:

> Yes, please set a final date as to when all usable items are due. I assume that an early date would move some of the 10 teachers in the "still working" category to the "dropped out" category. Those kits go out next week. However, our first training on a kit for which you're developing tests is "Magnetism and Electricity"

on September 20, then steady on after that. So, if we had pilot versions of tests on that date, we could administer them at training Sept. 20 and give them to trainees to take with them to administer to their students. These teachers would then put the completed tests in the kits when they return ship them. So, will it be possible to set a date when usable items are due such that you give yourself time to get pilot versions of the tests ready by Sept. 20?

By August 31, 2005—the cut-off date for final items, of the ten teacher-item writers who were "still working" on their items as of August 22, 2005, four had submitted final items; the remaining six had dropped out. Table 19 provides the final status of item writing as of August 31, 2005.

Table 19. Status of item writing as of August 31, 2005

Item Writer	**Items written for:**	**Status of "usable" items as of 8/19/05**	**8/31/05 update of "still working on"**
IW02	Gr 3: Human Body	Received 8/8/05	
IW04	Gr 3: Plant Growth & Development	Received 8/3/05	
IW13	Gr 3: Magnetism & Electricity	Received 7/20/05	
IW14	Gr 4: Magnetism & Electricity	Received 8/8/05/05	
IW03	Gr 3: Plant Growth & Development	Still working on	Dropped out 8/31/05
IW07	Gr 3: Plant Growth & Development	Still working on (a potential drop out since no items received)	Dropped out 8/31/05
IW10	Gr 3: Soils	Still working on	Dropped out 8/31/05
IW11	Gr 4: Good Chemistry	Still working on	Received 8/29/05
IW12	Gr 4: Magnetism & Electricity	Still working on	Dropped out 8/24/05
IW17	Gr 5: Landforms	Still working on	Received 8/31/05
IW18	Gr 5: Landforms	Still working on	Received 8/31/05
IW19	Gr 5: Motion & Design	Still working on	Received 8/25/05
IW20	Gr 5: Motion & Design	Still working on	Dropped out 8/25/05
IW15	Gr 4: Magnetism & Electricity	Still working on (?--25 initial items sent with no	Dropped out 8/31/05

Item Writer	Items written for:	Status of "usable" items as of 8/19/05	8/31/05 update of "still working on"
		specification sheets)	
IW05	Gr 3: Plant Growth & Development	Dropped out 8/2/05	
IW06	Gr 3: Plant Growth & Development	Dropped out 8/1/05	
IW08	Gr 3: Plant Growth & Development	Dropped out 7/28/05	
IW01	Gr 3: Human Body	Dropped out/no responses	
IW09	Gr 3: Soils	Dropped out/no responses	
IW12	Gr 4: Magnetism & Electricity	Dropped out 8/19	

From the 20 item writing workshop attendees, 8 (i.e., 40 percent of the initial item writers) actually submitted the requisite 16 items—8 teacher test items and 8 student test items—for which each teacher was paid $300 and received 0.5 renewal credits.

Factors that affected item generation and revision.

Item generation and revision were affected primarily by the TASC project participants, i.e., the TASC project director and the teacher-item writers. The TASC project director made various assumptions throughout the item writing part of phase 3. First, the TASC project director assumed that teachers, once they had been trained on a curriculum unit, would use that unit in their classrooms and therefore, would know the science content of the unit well enough to write test questions. TASC also assumed that the teachers, after receiving one half-day of item writing training, would be able to write items that were of "medium to hard" difficulty and required higher level thinking skills. Last of all, TASC assumed that compensation and renewal credits would be sufficient incentives for teachers to complete the item writing task.

In an August 24, 2006 interview of the TASC project director, this researcher asked what his initial expectations had been regarding the selection of teacher-item writers. He stated:

> … I thought we'd be able to find … a reasonable number, like 20, 25, teachers for over a couple kits that would … know the content well enough and could write higher order thinking skills items and so I did expect they'd be able to do it and … I was pretty surprised … I expected that they would be willing if paid to spend you know several days of work on it and that they'd be rigorous and … do their best.

Based on his statement, the TASC project director assumed that 1) the teachers would know the content well enough to write items, 2) the teachers would know how to write higher order thinking skill items, and 3) the teachers, if compensated, would be willing to write items for tests that they and their students would take.

TASC learned first-hand that the teachers, who had been trained and had used the unit in their classroom, frequently did not know the science content well enough to write items. In addition, there was no way of knowing how much of the curriculum unit the teacher actually used. Even though TASC included in each unit an informal self-report page, frequently the page was left blank by the classroom teacher. The amount of usage of a unit by a trained teacher was also not verified by TASC.

Teachers' lack of content knowledge surprised some of the TASC personnel. For example, in this researcher's August 24, 2006 interview of the TASC training director, he stated:

> I think what we .. hoped for was that … we would help teachers understand the content better and therefore they would help their students understand the content better. … we didn't realize how weak that content piece was for the teachers. That they knew so little.
>
> We didn't anticipate that they knew so little, maybe we should have but we didn't know. … the other piece was that we wanted to teach the teachers how to use these materials. And I think the most important goal was … that we wanted to teach the teachers how to use inquiry in the classroom. ... We'll use science kits as a way to get—inquiry-based science kits—as a way to get the teachers to do inquiry in the classroom … even if they are not comfortable with it themselves, we're going to help them try to get comfortable with the kits and then they'll e doing inquiry because they'll be doing the kits.
>
> … Kits were a means to that end and content was something that we had to do some but we didn't realize how much and what we have now realizes is that content information is … a key part of being able to do inquiry well.

Other TASC personnel had a more realistic view of teachers' science content knowledge and exactly what they expected the teachers to get out of the one-day, or two-day (separated by three weeks), TASC training. In this researcher's August 24, 2006, interview of the TASC curriculum director, he stated that while TASC's "goal over time was … to change … the way they [the teachers] teach, … I think … the real goal is can we get the people comfortable enough to feel like they can open up the box and start to use it with their kids and start down the long road to improving the way that they teach." This sentiment was reiterated in this researcher's August 24, 2006, interview of the TASC materials manager, who stated a similar goal of his training:

> the goal of my training is to get the teacher comfortable and willing to open the kit and use it with their students. And that is number one on the agenda for me and along with that is to teach them some science but it's to teach them enough so they feel comfortable doing it with the students.

In response to the question as to whether he assumed that the teachers would already have science content knowledge, the TASC materials manager responded:

> No. I assumed that … many would come with limited [science content knowledge] and my assumption from what I've seen in the classrooms is that if I can get them to [be] comfortable … with the science that's involved with the kit that they will be willing to open the box and try it out. … I had no dreams or aspirations of teachers walk[ing] out of my workshops knowing 20 or 30 percent more microbiology than when they came in.

Likewise, the TASC curriculum director was not surprised by the teachers' lack of science knowledge, stating:

> They don't understand the kits … and they don't know very much science. So I think in part when they work with us, they're still trying themselves to learn the science that's in that kit and to be able to see a question that's beyond, you know, what color did it turn when you added the drops, you have to know more.

Based on these statements, clearly some of the TASC personnel understood that the teachers did not have the science content knowledge prior to attending a one or two-day workshop and that the teachers were not going to gain sufficient science content knowledge in that short period of time to reflect a 20 percent increase in content knowledge, as proposed in TASC's evaluation plan. It appears, therefore, that TASC's expectation that teachers, who had received one or two days of TASC training and who may have used the units—to what extent was unclear—in their classrooms, write the items for these tests was overly optimistic.

The second assumption TASC made—that teachers, after attending a half-day item writing workshop, would be able to write a total of 16 “medium to hard” items that

required higher level thinking skills within three weeks—also appeared to be unrealistic. Based on the comments of a few of the item writers, even they were surprised by the difficulty of the task.

IW20 wrote in a July 10, 2005 email:

> Wow! This was more difficult than I thought it would be.

IW16 wrote in a July 11, 2005 email:

> … as I began to work on the questions I realized how difficult it was to use only one content goal at once, not have it knowledge based, and have the answers match those goals provided [in the test blueprint].

IW14 wrote in an August 17, 2005 email:

> I spent more time than I ever imagined, but it was a great experience. I have a new respect for people who create our test item questions for the end of grade tests.

The TASC project director thought that the teachers who were recruited from the TASC Institute Training (in June 2005) as item writers "would be the cream of the crop." What surprised him was "just … how ill-equipped teachers are to write higher-order thinking skills items. In fact, how ill-equipped they are to ask questions even of … their students … that require them to think."

The TASC curriculum director, however, did not think that having teachers write the test items would work. When asked why, he responded:

> Well, because the teachers are really just getting started at using this kind of teaching and … it's really different, and I think … it was beyond … for the broad range of teachers who come to the workshops … their capabilities to sort of see

> beyond how the kits really worked and what their teaching is like with these kits and how do you evaluate something like that; you know, it's just going to sort of be pretty cut and dried textbooky kind of test, not much thinking. And even then the questions would be very good.

He believed the questions would not be very good because the teachers did not have adequate science content knowledge.

TASC's assumption that compensation and renewal credits would be sufficient incentives for teachers to complete the item writing task also turned out to be unrealistic. As presented earlier in this section, 20 teachers attended the item writing workshops. Of those 20 teachers, 8 completed the task of submitting 16 "usable" items.

The process of reviewing items, requesting revisions, reviewing revisions, requesting additional revisions, reviewing re-revisions, etc., was laborious and time-consuming. Initially three to four weeks had been allotted for the completion of the item-writing task. Instead, the process took eight weeks to complete, right up to the August 31, 2005 cut-off date for final items. Factors that contributed to the amount of time it took to complete the item writing included teachers' inadequate science content knowledge and item writing skills.

One way inadequate science content knowledge manifested itself was through inaccurate item content. One such example was a question (Figure 8, below) submitted with an inaccurate depiction of a parallel circuit.

***Teacher* Science Item Specification Sheet for:**

☐ Grade 3 ☑ Grade 4 ☐ Grade 5

Competency Goal:	**Instructional Objective:**	**NC Thinking Skill(s):**
☐ Goal 1 ☐ Goal 2 ☑ Goal 3 ☐ Goal 4	3.08	☐ Knowledge ☑ Generating ☐ Organizing ☐ Integrating ☑ Applying ☐ Evaluating ☐ Analyzing
Difficulty Level: ☐ Easy ☑ Medium ☐ Hard	**Item Writer:**	**Artwork required?** ☐ Yes (if checked, please attach and document source) ☐ No

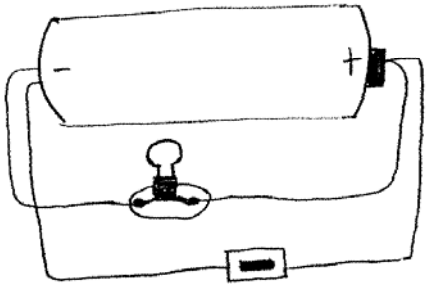
Science Test Item:

A group of students designed a circuit like the one below. They wanted to be able to light a bulb and ring a door bell using a single D-cell battery. They knew they could do this because:

A. The D-cell can only have a single pathway for electricity to travel through.

B. The D-cell can run several components if connected in a parallel circuit.

C. The D-cell cannot have only a single set of wires connected to it at one time.



Correct Answer: B

General Guidelines:

- ☐ Focus directly on the objective.
- ☐ Write stem as a complete statement/question.
- ☐ Write distractors of equal length.
- ☐ Write distractors using same context and similar ideas.
- ☐ Make distractors grammatically consistent with the stem.
- ☐ Make each distractor plausible (and document reasoning).
- ☐ Avoid using negatives in distractors.
- ☐ Check punctuation, spelling, and grammatical structure of item.
- ☐ Use artwork as needed (and document source(s)).
- ☐ Practice fair representation in sex and race, avoiding culture-specific references.

Figure 8. Initial multiple choice item from item writer 14.

Inadequate science content knowledge also manifested itself through the use of curriculum-specific language and examples. That is, the teachers did not have a thorough understanding of their content to generalize outside of the specific context of the curriculum unit. One such example (Figure 9, below) was submitted by IW20 who used a diagram that was specific to the curriculum unit. This researcher had a very difficult time trying to make item writers understand that the questions being developed were to evaluate grade-specific instructional objectives (e.g., NC SCS Grade 5, 4.05) and should not be written in such a way that they gave unfair advantage to those who had been exposed to the particular curriculum units.

***Teacher* Science Item Specification Sheet for:**

☐ Grade 3 ☐ Grade 4 X Grade 5

Competency Goal:

☐ Goal 1
☐ Goal 2
☐ Goal 3
☒ Goal 4

Instructional Objective:

____Determine factors that affect motion.

4.05

NC Thinking Skill(s):

☐ Knowledge ☐ Generating
☐ Organizing ☐ Integrating
☐ Applying ☐ Evaluating
☒ Analyzing

Difficulty Level:

☐ Easy
☒ Medium
☐ Hard

Item Writer:

IW20

Artwork required?

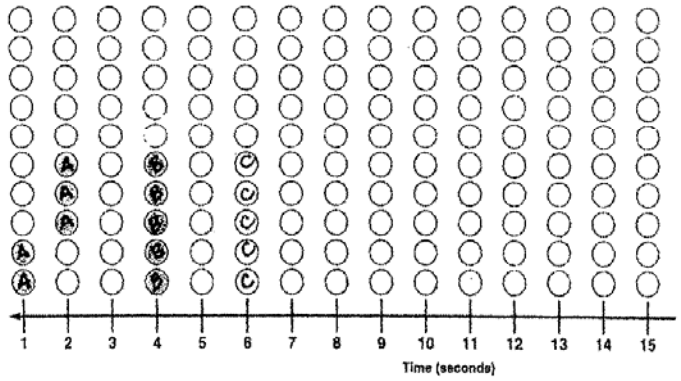
☒ Yes (if checked, please attach and documen source) STC Motion and Design manual Recording Sheet 4-A

☐ No

Science Test Item:

Graphing Data: How Load Affects the Time a Vehicle Travels

Number of washers we will use: 16



A= vehicle only B= vehicle + 1 block C= vehicle + 2 blocks

What can you conclude about the effects of load on a vehicle's motion?

a. The heavier the vehicle the longer it takes to respond to a force.
b. The lighter the vehicle the longer it takes to respond to a force.
c. The load has no bearing on the response to force.

Correct Answer: _____a_____

Figure 9. Example of curriculum-specific question

Ambiguous stems and implausible distractors were other examples of inadequate science content knowledge. Figure 10 provides an example of an ambiguous stem in a question submitted by IW19.

Think about what happens when you add tires to the wheels of a vehicle. What would happen to the friction between the road and the vehicle when the tires are added?
A) It will increase the friction.
B) It will decrease the friction.
C) It will not change the friction.

Correct Answer: ___A_____

Figure 10. Example of ambiguous stem.

Implausible distractors appeared in many forms. For example, IW08 submitted a question that read, "This picture show [sic] a bee …" where one of the four distractors was "giving color to a flower". Another example of implausible distractors is given in Figure 11, a question submitted by IW06.

The heart is similar to leaves on a plant because it is
a) in the center of the body
b) shaped like a leaf
c) carries blood through the body
d) helpful to human's balance

Correct Answer: ___c_____

Figure 11. Example of implausible distractors

Teachers inadequate item writing skills were another factor that affected the item writing task of phase 3 contributing to the prolonged period of time allotted for item writing. Initially, teachers submitted knowledge-level questions, even though this category was not included on any of the test blueprints. In a June 28, 2005 email sent to all the item writers, this researcher stated:

> … as I stated at the workshops, the test blueprints are our guides for item writing and none of the blueprints include the Knowledge-level NC Thinking Skill. The reason for this is that we are interested in determining whether students/teachers know how to *use* science content/process skills.

Even after this reminder, item writers submitted knowledge-level questions, yet identified them as a higher-level question. One such example is provided in Figure 12, an item submitted by IW11. This item writer indicated the item required the "analyzing" and "evaluating" NC Thinking Skills, even though the item did not require recognizing and articulating parts that constitute a larger whole (analyzing) or judging the quality, credibility, worth, and/or practicality of ideas (evaluating) as stated in Appendix A of the *Multiple Choice Item Writing Workbook.* This practice of misidentification of NC Thinking Skills may not be uncommon as this researcher found many such occurrences in her recent review of draft benchmark items written by teachers in a local NC school district.

Teacher Science Item Specification Sheet for:

☐ Grade 3 ☒ Grade 4 ☐ Grade 5

Competency Goal:	Instructional Objective:	NC Thinking Skill(s):
☐ Goal 1 ☐ Goal 2 ☐ Goal 3 ☒ Goal 4	4.03	☐ Knowledge ☐ Generating ☐ Organizing ☐ Integrating ☐ Applying ☒ Evaluating ☒ Analyzing
Difficulty Level:	**Item Writer:**	**Artwork required?**
☐ Easy ☒ Medium ☐ Hard	IW 11	☐ Yes (if checked, please attach and document source) ☒ No

Science Test Item:

What vitamin is this person probably lacking?

A. Vitamin C
B. Vitamin B
C. Vitamin D
D. Vitamin K

Correct Answer: A

General Guidelines:

- ☐ Focus directly on the objective.
- ☐ Write stem as a complete statement/question.
- ☐ Write distractors of equal length.
- ☐ Write distractors using same context and similar ideas.
- ☐ Make distractors grammatically consistent with the stem.
- ☐ Make each distractor plausible (and document reasoning).
- ☐ Avoid using negatives in distractors.
- ☐ Check punctuation, spelling, and grammatical structure of item.
- ☐ Use artwork as needed (and document source(s)).
- ☐ Practice fair representation in sex and race, avoiding culture-specific references.

Figure 12. Example of misidentified NC Thinking Skill item

Even though general guidelines for item writing were included in the *Multiple Choice Item Writing Workbook* and on the science item specification sheets used by item writers to submit their items, this researcher reviewed items:

- that did not focus directly on the objective;
- with incomplete statements and/or questions in the stem;
- with distractors of unequal length, dissimilar context, and/or grammatically inconsistent with the stem; and
- with incorrect punctuation, spelling, and grammatical structure.

To summarize, the item writing task of phase 3 was predominantly affected by the TASC project participants, i.e., the TASC project director and the teacher-item writers. The TASC project director made various assumptions upon which the item writing task was based. The project director assumed that:

1) teachers, once they had been trained on a curriculum unit, would use that unit in their classrooms and therefore, would know the science content of the unit well enough to write test questions;
2) teachers, after receiving one half-day of item writing training, would be able to write items that were of "medium to hard" difficulty and required higher level thinking skills; and
3) compensation and renewal credits would be sufficient incentives for teachers to complete the item writing task.

These assumptions were shown to be overly optimistic. Teachers did not know the content well enough to write items; teachers, after taking a half-day item writing

workshop, struggled to write higher order thinking skills items; and compensation plus renewal credits did not prevent 60 percent of the item writers from dropping out.

The teachers' inadequate science content knowledge and item writing skills also affected the item writing task. Examples of inadequate science content knowledge included inaccurate item content, items based specifically on the curriculum units, ambiguous item stems, and implausible distractors. Examples of inadequate item writing skills included the frequent submission of knowledge-level questions; the misidentification of items as higher order thinking skill items when, in fact, the items were knowledge-level items; and poorly constructed items.

The discussion now moves to the assembly of the pilot tests.

Pilot Test Assembly

In the TASC project director's August 22, 2005 email, mentioned in the previous subsection, he had indicated that September 20 was the first TASC training date on a curriculum unit for which CERE was developing a test. In addition, he had stated:

> So, if we had pilot versions of tests on that date, we could administer them at training Sept. 20 and give them to trainees to take with them to administer to their students. These teachers would then put the completed tests in the kits when they return ship them.

To communicate her understanding of CERE's and TASC's remaining responsibilities under the subcontract, this researcher summarized the tasks to be completed by CERE and by TASC in her August 26, 2005 email to the TASC project director, which stated:

> After our conversation, I spoke with [the CERE acting director] about what you proposed concerning the tests we are developing for TASC. To summarize, I will complete student and teacher tests for:
>
> - Grade 3 Plant Growth & Development
> - Grade 3 Human Body
> - Grade 3 Soils (tentative)
> - Grade 4 Magnetism & Electricity
> - Grade 4 Food Chemistry
> - Grade 5 Landforms
> - Grade 5 Motion & Design
>
> You will request permission from the publishers of Grade 5 Ecosystems, Grade 8 Earth History, and Grade 8 Micro-Life to use the test items in their materials. [The CERE acting director] reminded me that we (CERE/UNCG) will need that permission to be in writing. That way, both TASC and CERE/UNCG are protected from any potential problems.
>
> Teacher pre-tests will be administered by TASC personnel prior to the teachers' TASC training. The teachers will receive their TASC training and, at the end of the training, they will be given pre- and post-tests for the students. The teachers will administer the student pre-tests before they begin to use the TASC science curriculum units. After the science unit is completed (after about nine or ten weeks), the teacher will administer the student post-tests. One question I forgot to ask you is who will administer the teacher post-tests? Another question is, can we (CERE/UNCG) receive the pretest results as soon as the pre-testing is completed?
>
> In addition to written permission from the publishers to use their items, [the CERE acting director] also brought up another issue. How does TASC want us (CERE/UNCG) to proceed when we recognize that an item (or items) from the publishers are poor items, based on item writing guidelines?

This communication makes clear that, even though there were no teacher-item writers for the Grade 3 Soils, Grade 5 Ecosystems, Grade 8 Earth History, and Grade 8 MicroLife tests, TASC wanted these tests developed and became responsible for helping to provide the requisite items.

In his emailed response, the TASC project director restated his understanding of TASC's responsibilities:

> We'll get the permissions in writing to use items from FOSS (Lawrence Hall of Science) and STC (NSRC). We will administer the pre-tests to teachers who attend training at GSK, and we can send you the completed pretests the day after teachers complete them here at GSK (or if it would be OK, we could collect a week's worth and send them on Thursdays). Teachers will NOT receive the pre-posts for students and the teacher post-tests at the end of training, as you suggested. The pre-posts for students and the post tests for teachers will be packed in the kits. We'll ask teachers to administer the student pre-test before they begin the unit. We'll also ask them to administer the student post-test after the 9 weeks, then put the completed tests in the kits to ship back to us. We plan to ask teachers to self-administer the teacher post-tests and put them in the kits with their student tests. We will ask them to identify the test with the last 4 digits of their SS#'s, and not to make up any numbers. Regarding use of test items from the kit teacher's guides, TASC wants CERE/UNC-G not to use any items from publishers that you deem to be poor, based on your item writing guidelines.

Also on August 26, 2005, this researcher, using the items she had accepted from IW02 and IW04, assembled draft student and teacher tests for Grade 3 Human Body (HB) and Grade 3 Plant Growth & Development (PGD), respectively.

By the end of August 2005, it became obvious to this researcher and the CERE acting director that, with the current TASC-CERE subcontract due to expire by September 30, 2005, an extension would be necessary in order for CERE to assemble, pilot, and analyze the 11 tests TASC expected CERE to complete. Therefore, on August 29, 2005, the CERE acting director emailed the TASC project director requesting a letter of extension from TASC because "we got such a late start" that the current subagreement would need to be extended beyond its September 30, 2005, deadline in order to complete the work on the tests.

On August 30, 2005, the TASC program director responded to the CERE acting director, who had submitted a proposed budget to him for his review, with his understanding of the work to be completed under the (extended or new) TASC-CERE subcontract. He stated:

> I can't find your revised budget nor any email from you or Terry [this researcher] with that budget. I have only the one [you] sent to me . . . on August 1. I'm sure you sent the revised one, but it didn't come through. Could you send it again? Sorry. That budget would cover the following:
>
> CERE will complete student and teacher tests for:
>
> - Grade 3 Plant Growth & Development
> - Grade 3 Human Body
> - Grade 3 Soils (tentative)
> - Grade 4 Magnetism & Electricity
> - Grade 4 Food Chemistry
> - Grade 5 Landforms
> - Grade 5 Motion & Design
>
> As I discussed with Terry, we will refocus our testing on just 3 grade levels: grades 3, 5, and 8. Note that 2 grade 4 kits are listed above. I do not want to waste the work already put into these, so, until it proves too costly to continue with the grade 4, I think we should go ahead and administer Manetism [sic] & Electricity and Food Chemistry. I'm not dead set on this, so please advise. If it will help, I can send you FOSS's 34 Magnetism & Electricity test items, many of which look quite good to me.
>
> I will request permission from the publishers of the following to use the test items in their materials in writing, and TASC staff will send you excerpts of these by the end of the week:
>
> - Grade 5 Ecosystems STC
> - Grade 8 Earth History FOSS-Delta (we already have this)
> - Grade 8 Micro-Life SEPUP
>
> To round out the test items for grades 3, 5, and 8, TASC staff will supply 15 items each for the following by the end of the week, for you to revise or recommend revisions:
>
> - Grade 3 Investigating Objects in the Sky (T.R.A.C.S.) BSCS
> - Grade 5 Ecosystems (STC)

- Grade 5 Investigating Weather Systems (T.R.A.C.S.) BSCS

In addition, as needed, TASC staff will work with you to produce enough items for the "Soils" test, grade 3.

You have the written permission from FOSS. As I understand it, CERE will look all of the items over for any glaring problems. CERE will then test the commercially-produced tests (listed above) and the items TASC staff will send to see how the items behave. The items will be tested in the first cycle this year. The tests must be ready to hand out on Sept. 20. To do that, TASC needs the tests a week before Sept. 20 to prepare them. Once we have pre- and post-results, CERE will analyze the tests to prepare them [to] use in evaluation. If it helps with your budget, TASC has some capability to put responses on Scantron sheets and program the scanning software. In that case, we could send you the sheets and the raw data for your analysis.

TASC personnel will administer teacher pre-tests to teachers at TASC training. CERE/UNCG will receive pretest results the week pre-testing is completed. After the teachers complete using the unit with their students in about 9 weeks, they will self-administer the post-test. Kits will contain both pre-tests and post-tests for the students. Teachers will administer student pre-tests before they begin to use the curriculum units and administer the student post-tests after the science unit is completed about nine weeks later.

Please let me know if this meets with yours and Terry's understanding of our discussions so far, and help me fill in anything I have missed. Also, if any of these ideas are off base, please let me know.

A summary of all the tests, by kit and grade level, to be provided from the sources listed above and analyzed by CERE is as follows:

- Grade 3 Plant Growth & Development
- Grade 3 Human Body
- Grade 3 Soils (tentative)
- Grade 4 Magnetism & Electricity (under discussion)
- Grade 4 Food Chemistry (under discussion)
- Grade 5 Landforms
- Grade 5 Motion & Design
- Grade 5 Ecosystems STC
- Grade 5 Investigating Weather Systems (T.R.A.C.S.) BSCS
- Grade 8 Earth History FOSS-Delta (we already have this)
- Grade 8 Micro-Life SEPUP

To summarize, 11 tests (teacher and student versions) were to be assembled—even though items had not yet been written and/or reviewed for five of the tests—and ready to be piloted in TASC's first training cycle in the 2005-2006 academic year that would begin September 7, 2005. By August 31, 2005, "usable" items had been accepted by CERE from teacher-item writers for the following tests:

- Grade 3 Human Body (IW02)
- Grade 3 Plant Growth & Development (IW04)
- Grade 4 Food Chemistry (IW11)
- Grade 4 Magnetism & Electricity (IW13 and IW14)
- Grade 5 Landforms (IW17 and IW18)
- Grade 5 Motion & Design (IW19)

The tests for which TASC would provide items, either through obtained permission or through writing the items, were:

- Grade 3 Investigating Objects in the Sky
- Grade 3 Soils
- Grade 5 Ecosystems
- Grade 5 Investigating Weather Systems
- Grade 8 Earth History
- Grade 8 MicroLife

On August 31, 2005, the TASC project director emailed Duke's Office of Research Support, requesting guidance as to how to proceed—extend the current subaward or issue a new one—in order that CERE might continue its work. He then emailed the CERE acting director with a summary of the response he had received from the Office of Research Support:

> He would like for us to terminate the '04-'05 subaward with you on Sept. 30 and then re-issue you a new '05-'06 subaward that begins Oct. 1. The new subaward will include a budget containing all of the unexpended funds from the '04-'05 subaward (unspent by Sept. 30). I showed him your budget, but he needs to have that approved by … UNC-G before he can do anything with it. So, please forward that budget …. When your last invoice has cleared our '04-'05 budget, we will take the amount remaining and add it to the budget you sent me, which has been cleared by [UNC-G].
>
> This would not involve any letters extending the previous subaward.

Having established that a new subcontract would be issued by Duke-TASC to UNCG-CERE, this researcher immediately began the test assembly process.

Figure 13 presents TASC's 2005-2006 training schedule. The fall cycle's Session 1 dates were the dates on which TASC wanted to administer the pilots of the teacher pretests (five of which had yet to be developed) to their workshop attendees. In addition, TASC wanted the pilots of the student pretests, student posttests, and teacher posttests assembled, printed, and delivered to the TASC Resource Center (i.e., warehouse) in Durham in time to be included in the applicable curriculum units that shipped out within two or three days of Session 1.

TASC Training Schedule at GlaxoSmithKline, 2005-2006

		Fall Cycle Training dates		Winter Cycle Training Dates	
		Session 1	Session 2	Session 1	Session 2
K	Wood & Paper	9/7/05	9/28/05	not offered	
K	Comparing & Measuring	9/12/05	--	not offered	
K	Ant Homes Under the Ground	9/8/05	--	2/2/06	--
1st	Solids & Liquids	not offered		2/16/06	3/9/06
1st	Balance & Motion	9/12/05	--	1/31/06	--
1st	Organisms	9/15/05	10/6/05	2/9/06	3/2/06
1st	Pebbles, Sand & Silt	9/13/05	10/4/05	2/7/06	2/28/06
2nd	Changes	9/22/05	10/13/05	not offered	
2nd	Lifecycle of Butterflies	9/15/05	--	2/9/06	--
2nd	Sound	9/14/05	10/5/05	1/31/06	2/21/06
2nd	Air & Weather	9/8/05	9/29/05	2/1/06	2/22/06
3rd	Soils	9/7/05	9/28/05	not offered	
3rd	Plant Growth & Development	9/13/05	10/4/05	2/7/06	2/28/06
3rd	Human Body	9/22/05	10/13/05	2/8/06	3/1/06
3rd	Investigating Objects in the Sky	9/7/05	--	1/31/06	--
4th	Magnetism & Electricity	9/20/05	10/11/05	2/14/06	3/8/06
4th	Food Chemistry	not offered		2/8/06	3/1/06
4th	Rocks & Minerals	not offered		2/1/06	2/22/06
4th	Animal Studies	9/8/05	9/29/05	2/2/06	2/23/06
5th	Investigating Weather Systems	9/14/05	10/5/05	2/2/06	2/23/06
5th	Landforms	9/13/05	10/4/05	2/7/06	2/28/06
5th	Motion & Design	not offered		2/14/06	3/7/06
5th	Ecosystems	9/15/05	10/6/05	2/9/06	3/2/06
6th	Solar System	9/20/05	--	not offered	
6th	Energy Transfer & Transformation	not offered		2/8/06	--
6th	Cycling of Matter & Pop. Dynamics	11/1/05	--	not offered	
6th	Earth's Crust	not offered		4/4/06	--
7th	Thrill Ride	not offered		2/15/06	3/8/06
7th	Weather and Water	9/21/05	10/12/05	2/16/06	3/9/06
8th	Earth History	9/21/05	10/12/05	not offered	
8th	Micro-Life	9/20/05	10/11/05	2/15/06	3/8/06

Figure 13. TASC 2005-2006 Training Schedule

Table 20, the expected pattern for pilot test assembly task, shows a straightforward process of assembling pilot tests from items provided—either by the final teacher-item writers or by TASC—resulting in a teacher set and a student set of four grade 3 tests, two grade 4 tests, four grade 5 tests, and two grade 8 tests. As tables 21 through 31 will show, the process was anything but straightforward.

Table 20. Expected pattern for pilot test assembly task

Resources	Activities	Outputs	Outcomes
• Accepted items from the 8 teacher item-writers • Items received from TASC • TASC scientists • Test developer	• Test developer: assemble tests from items accepted from teacher-item writers and from TASC • TASC scientists: review tests for content	• Draft teacher tests • Draft student tests	• Grade 3 tests: o Human Body o Investigating Objects in the Sky o Plant Growth & Development o Soils • Grade 4 tests: o Food Chemistry o Magnetism & Electricity • Grade 5 tests: o Ecosystems o Investigating Weather Systems o Landforms o Motion & Design • Grade 8 tests: o Earth History o MicroLife

Table 21 presents the pilot tests' delivery schedule. Keeping in mind that the assembly process did not begin until the very end of August 2005, the table illustrates the severe time constraints under which the creation, assembly, packaging, and delivery of the tests took place. Tables 22 through 31 document the test assembly for each test. The tables are arranged in the order of each test's targeted pilot date (i.e., Session 1 training date).

Table 21. Pilot tests' delivery dates

Pilot Tests:	Session I TASC Workshop Training Date	Date of 1st Draft of Pilot Test	Date of Final Version of Pilot Test	Tchr Pre-tests	Science Kits Shipped Out by TASC
Gr 3 Soils[1, 2]	9/7/2005	9/3/2005	n/a (workshop cancelled)	n/a	n/a
Gr 3 Investigating Objects in the Sky[2]	9/7/2005	9/3/2005 (from TASC)	9/6/2005	17	9/9/2005
Gr 3 Plant Growth & Development	9/13/2005	8/31/2005 (to TASC)	9/12/2005	10	9/15/2005
Gr 5 Landforms	9/13/2005	9/7/2005 (to TASC)	9/12/2005	19	9/15/2005
Gr 5 Investigating Weather Systems[2]	9/14/2005	9/3/2005 (from TASC)	9/13/2005	8	9/16/2005
Gr 5 Ecosystems[2]	9/15/2005	9/3/2005 (from TASC)	9/12/2005	11	9/16/2005
Gr 4 Magnetism & Electricity	9/20/2005	n/a--test dropped by TASC on 9/14/2005	n/a	n/a	9/23/2005
Gr 8 Micro-Life[2,3]	9/20/2005	9/3/2005 (from TASC)	10/10/2005	20	9/23/2005
Gr 8 Earth History[2]	9/21/2005	9/3/2005 (from TASC)	9/19/2005	9	9/23/2005
Gr 3 Human Body	9/22/2005	8/31/2005 (to TASC on 9/13/2005)	9/16/2005	28	9/23/2005
Gr 4 Food Chemistry[4]	2/8/06	n/a	n/a	n/a	n/a
Gr 5 Motion & Design--Forms A and B	2/14/06	11/22/05 (to TASC)	2/6/06	17	2/17/06
Gr 8 Micro-Life	2/15/06	9/3/05	10/10/2005	19	2/17/06

[1]two item writers dropped out; all items provided, or written by, TASC
[2]workshop cancelled due to low registration

Pilot Tests:	Session I TASC Workshop Training Date	Date of 1st Draft of Pilot Test	Date of Final Version of Pilot Test	Tchr Pre-tests	Science Kits Shipped Out by TASC

[3]missed 9/20/2005 date; TASC wanted to administer pilot test at Session 2 workshop on 10/10/2005

[4]Items had been accepted from two teacher-item writers; however, this test was dropped from the 2005-2006 CERE-TASC subcontract.

Table 22. Grade 3 *Soils* pilot test assembly

Grade 3 Soils pilot test assembly	
8/31/05	IW09 and IW10 dropped out as item writers.
9/4/05	TASC project director (p/d) sent 15 Soils test items.
9/5/05	TB sent draft of test to TASC p/d with questions on some of the items.
9/6/05	1. TB received from TASC p/d revised Soils test with input from TASC Soils workshop scientist. 2. TB sent by delivery service Scantron sheets for test to TASC. 3. Soils workshop cancelled by TASC due to lack of registrants.
9/7/05	Session I of two-day workshop (Fall Cycle)—cancelled.

Table 23. Grade 3 *Investigating Objects in the Sky (IOS)* pilot test assembly

Grade 3 Investigating Objects in the Sky (IOS) pilot test assembly	
9/3/05	1. TASC p/d emailed to me "tests … from TASC staff". 2. TB provided feedback and included "additional items for [his] consideration".
9/5/05	1. TASC training director (t/d) responded to 9/3 email TB sent to TASC p/d. 2. TB emailed to TASC p/d revised teacher and student tests (090505 Gr 3 IOS), with a question on item 8 and with items' choices arranged

Grade 3 Investigating Objects in the Sky (IOS) pilot test assembly	
	alphabetically. 3. TASC communicated with TB telephonically—answering her question on item 8 and indicating there would be *one* version of the test for both teachers and students. 4. TB corrected item 8 and emailed revised IOS tests with "only difference between the two tests" their background information questions (i.e., Section 2).
9/6/05	1. TASC p/d emailed the test, making a minor revision to test (boxes around one of the diagrams) and changing the formatting of the test. 2. TB re-revised the test sent by TASC p/d, correcting his (re)formatting (and asking him in the future to refrain from revising the test formatting. 3. TB emailed to TASC p/d instructions to be read to the teachers by the TASC test administrator (i.e., TASC scientist who would be presenting IOS workshop).
9/7/05	1. Session I of two-day workshop (Fall Cycle): TASC made copies of the Gr 3 IOS teacher pretest for the 17 workshop registrants. 2. TASC p/d emailed TB with instructions for color-coding student tests the same as the teacher tests (i.e., pretest – white; posttest – blue) and requesting tests be shipped to TASC warehouse. He stated: The ship date on that kit is tomorrow. We will probably make a test packet containing all of these in an envelope and pack that in the kit with the test taking instructions. However, we'll wait on shipping it until we get the tests. All kits for which teachers receive training (no matter when) are shipped after the teachers attend session 1. 3. TB responded with preference that student tests be color coded differently than teachers (pretest – yellow, posttest – green) to facilitate separating them upon their return to CERE. TB also indicated tests would have to be delivered by Federal Express because UNCG had temporarily banned all travel. 4. TASC p/d acknowledged TB's email and indicated the 9/9/05 arrival date of the tests would work for TASC.
9/8/05	TB had delivered to TASC 17 teacher posttests, 460 (17 teachers x 27 students/teacher) student pretests and Scantrons, 460 student posttests and Scantrons, and pretest and posttest administration instructions for teachers to read to their students.
9/9/05	1. TASC t/d emailed to me a corrected answer key for IOS test. 2. TASC p/d emailed TB that tests had arrived at the TASC Resource

Grade 3 Investigating Objects in the Sky (IOS) pilot test assembly	
	Center (i.e., TASC warehouse). 3. Gr 3 IOS curriculum units shipped out of TASC Resource Center.

Table 24. Grade 3 *Plant Growth & Development (PGD)* pilot test assembly

Grade 3 Plant Growth & Development (PGD) pilot test assembly	
8/26/05	Using items written by IW04, TB assembled first draft of Gr 3 PGD test.
8/31/05	TB facsimiled to TASC p/d the first draft.
9/2/05	TASC p/d returned by facsimile "changes to the Plant Growth & Development test items that [TASC training director] (who teaches the Plant Growth & Dev. kit) suggests."
9/8/05	1. TB emailed to TASC p/d her feedback re: TASC's comments on Gr 3 PGD test 2. TASC p/d responded to TB that he "wasn't sure how to respond to your ideas, and forwarded it on to [TASC t/d] and he'll email you as soon as he can."
9/9/05	1. In an email to TASC p/d re: teacher pretests for week of 9/12/05, TB reminded him that she needed *at least* one day's lead time to get tests copied and then packed to be delivered to TASC in time for it to include them in the science kits being shipped out from the TASC warehouse. She stated: "This means that we must have the final version of next week's tests—Gr 3 Plant G&D, Gr 5 Landforms, and Gr 5 Ecosystems—to the print shop by Wed., 9/14, for them to be ready by Thur., 9/15, and delivered to [the TASC warehouse manager] by Fri., 9/16." 2. TASC t/d emailed to TB that his revisions had not been included in the 9/2 facsimile and that he was attaching them to the present email. He also indicated he would call regarding TB's additional questions. 3. TASC t/d emailed to TB "two additional questions that I guess are okay." 4. TASC p/d emailed to TB thanking her for the reminder about the copying and printing deadlines and providing the number of registrants for four of the workshops to be held 9/13, 9/14, and 9/15 (i.e., PGD, LDF, IWS, and ECO). 5. TB emailed to TASC t/d the assembled test that incorporated all his revisions and asking him to take a final review of the test, including

Grade 3 Plant Growth & Development (PGD) pilot test assembly	
	making a decision on whether one particular item was to be kept on the test or eliminated. 6. TASC t/d emailed TB with "several revisions." 7. TB responded to the TASC t/d's email with a PGD test that incorporated his most recent revisions; she sent a copy to the TASC p/d.
9/11/05	TASC t/d responded to TB's last 9/9 email: "This looks better. More suggestions …"
9/12/05	1. TASC p/d indicated the copy of the PGD test looked good to him and that once the TASC t/d approved it, it was ready to be printed. 2. TB emailed to the TASC t/d that she had incorporated his changes, and she responded to his question about a set of diagrams for one of the questions. 3. TB emailed a scanned image of the diagrams to the TASC t/d for his approval, which he provided telephonically. 4. TB emailed to TASC p/d and TASC t/d that the test had been sent to the printer, that it would be ready for pick-up on 9/13, and that she would have them delivered by Federal Express to the TASC warehouse for 9/14 delivery. She stated: "They will be 'classroom' packed; that is, one teacher post-test + 27 student pre-tests + 27 student post-tests = one classroom pack."
9/13/05	TB emailed (4:30 pm) to the TASC p/d that she had just picked up the copying of the two tests (Gr 3 PGD and Gr 5 LDF). She stated: "Since it would not be possible for me to single-handedly 'classroom pack' the shipment … in time for delivery tomorrow morning, the tests will not be delivered until Thursday."
9/13/05	Session I of two-day workshop (Fall Cycle): TASC made copies of the Gr 3 PGD teacher pretest for the 10 workshop registrants.
9/14/05	TB emailed to the TASC p/d that the "tests [270 student pretests, 270 student posttests, 10 teacher posttests] were delivered to TASC Resource Center around 4:30 pm this afternoon (Wed.)."
9/15/05	Gr 3 PGD curriculum unit shipped out of TASC Resource Center.

Table 25. Grade 5 *Landforms (LDF)* pilot test assembly

Grade 5 Landforms (LDF) pilot test assembly	
9/2//05	Using items written by IW17 and IW18, TB assembled draft LDF teacher and student tests.
9/7/05	TB emailed to TASC p/d first draft of LDF teacher pretest.
9/9/05	1. TB emailed TASC p/d asking if he had looked at the teacher pretest and reminding him that she needed *at least* one day's lead time to get tests copied and then packed to be delivered to TASC in time for it to include them in the science kits being shipped out from the TASC warehouse. She stated: "This means that we must have the final version of next week's tests—Gr 3 Plant G&D, Gr 5 Landforms, and Gr 5 Ecosystems—to the print shop by Wed., 9/14, for them to be ready by Thur., 9/15, and delivered to [the TASC warehouse manager] by Fri., 9/16." 2. TASC p/d responded to TB's email and indicated that the TASC t/d was "looking at the Gr 5 Landforms test, and I will send you his comments by the end of the day today." He also provided head counts for the upcoming week's TASC workshops (i.e., PGD, LDF, IWS, and ECO).
9/10 or 11/05	TASC t/d emailed his comments to TASC p/d.
9/12/05	1. TASC p/d emailed to TB the LDF comments from the TASC t/d. 2. TB emailed the LDF test with TASC t/d's revisions, and asked a question about the use of a "compass rose" on application level and analyzing level questions. 3. TASC t/d agreed to drop the use of the "compass rose". 4. TB emailed to TASC p/d and TASC t/d that the test had been sent to the printer, that it would be ready for pick-up on 9/13, and that she would have them delivered by Federal Express to the TASC warehouse for 9/14 delivery. She stated: "They will be 'classroom' packed [by CERE]; that is, one teacher post-test + 27 student pre-tests + 27 student post-tests = one classroom pack."
9/13/05	Session I of two-day workshop (Fall Cycle): TASC made copies of the Gr 5 LDF teacher pretest for the 19 workshop registrants.
9/14/05	Student pretests (513), student posttests (513), and teacher posttests (19) delivered to TASC Resource Center.
9/15/05	Gr 5 LDF curriculum unit shipped out of TASC Resource Center.

Grade 5 Landforms (LDF) pilot test assembly	
9/16/05	TASC t/d emailed to TB the LDF answer key.

Table 26. Grade 5 *Investigating Weather Systems (IWS)* pilot test assembly

Grade 5 Investigating Weather Systems (IWS) pilot test assembly	
9/3/05	TASC p/d emailed to me "tests … from TASC staff".
9/8/05	1. TB emailed to TASC p/d feedback re: IWS test. 2. TASC p/d responded to TB's comments with additional revisions.
9/9/05	1. TB acknowledged receipt of revisions and stated she would incorporate them and then return email the revised teacher test. 2. TASC p/d responded to TB's email and provided head counts for upcoming week's TASC workshops (i.e., PGD, LDF, IWS, and ECO). 3. TB emailed to TASC p/d revised IWS teacher pretest.
9/12/05	1. TASC p/d responded to TB's 9/9 email: "I've read over Inv. Weather Systems. It's OK as is, but I found 2 small tweaks I should have thought about earlier. …" 2. TB responded to TASC p/d's 9/12 email.
9/13/05	TB caught a formatting error on IWS teacher pretest, corrected it, and resent to TASC p/d the final version of the teacher and student IWS tests.
9/14/05	Session I of two-day workshop (Fall Cycle): TASC made copies of the Gr 5 IWS teacher pretest for the 8 workshop registrants.
9/16/05	1. Student pretests (216), student posttests (216), and teacher posttests (8) delivered to TASC Resource Center. 2. Gr 5 IWS curriculum unit shipped out of TASC Resource Center.
9/19/05	TASC p/d emailed to TB the IWS answer key.

Table 27. Grade 5 *Ecosystems (ECO)* pilot test assembly

Grade 5 Ecosystems (ECO) pilot test assembly	
9/3/05	TASC p/d emailed to me "tests … from TASC staff".
9/9/05	1. In an email to TASC p/d re: teacher pretests for week of 9/12/05, TB reminded him that she needed *at least* one day's lead time to get tests copied and then packed to be delivered to TASC in time for it to include them in the science kits being shipped out from the TASC warehouse. She stated: "This means that we must have the final version of next week's tests—Gr 3 Plant G&D, Gr 5 Landforms, and Gr 5 Ecosystems—to the print shop by Wed., 9/14, for them to be ready by Thur., 9/15, and delivered to [the TASC warehouse manager] by Fri., 9/16." 2. TASC p/d responded to TB's email and provided head counts for the upcoming week's TASC workshops (i.e., PGD, LDF, IWS, and ECO). 3. TB emailed to TASC p/d her feedback re: ECO test items. She had "a lot of questions about these test items" (e.g., what instructional objectives were being assessed, whether questions were the same yet worded differently, mismatch with test blueprint).
9/12/05	1. TASC t/d responds to TB's 9/9 email to TASC p/d, providing "a new version" of the ECO questions. 2. TB incorporated TASC t/d's changes and returned the revised ECO teacher pretest. 3. TASC t/d emails to TB "two simple corrections." He states that he had made the changes on the previously sent version.
9/15/05	Session I of two-day workshop (Fall Cycle): TASC made copies of the Gr 5 ECO teacher pretest for the 11 workshop registrants.
9/16/05	1. Student pretests (297), student posttests (297), and teacher posttests (11) delivered to TASC Resource Center. 2. TASC t/d emails to TB the IWS answer key. 3. Gr 5 ECO curriculum unit shipped out of TASC Resource Center.

Table 28. Grade 4 *Magnetism & Electricity* pilot test assembly

Grade 4 Magnetism & Electricity pilot test assembly	
8/31/05	CERE had accepted final items from IW13 and IW14.

Grade 4 Magnetism & Electricity pilot test assembly	
9/13/05	TB emailed TASC p/d that she "noticed on the TASC training calendar that there is a Gr 4 Magnetism & Electricity workshop on Tuesday, Sept. 20th. Are we doing tests for this unit?"
9/14/05	TASC p/d responded: "We are not writing tests for Magnetism & Electricity."
9/20/05	Session I of two-day workshop (Fall Cycle)
9/23/05	Gr 4 M&E curriculum units shipped out of TASC Resource Center.

Table 29. Grade 8 *MicroLife (ML)* and *Earth History (EH)* pilot tests assembly

Grade 8 MicroLife (ML) pilot test assembly and Grade 8 Earth History (EH) pilot test assembly*	
9/3/05	TASC p/d emailed to me "tests … from TASC staff".
9/4/05	EH: The TASC p/d emailed to TB that he had sent the wrong EH test on 9/3 and to ignore it and to review the one attached to the current email.
9/14/05	1. TB emailed to TASC p/d her feedback re: the Gr 8 tests (i.e., ML and EH) stating that her "overall impression of these two tests is that they need quite a bit of work." For instance: a. TASC had not yet provided to her the percentages of their instructional time spent on the NC SCS instructional objectives covered by the ML and EH workshops. b. Some of the questions provided by TASC were knowledge-level. c. The use of "all of the above" and/or "none of the above" were seldom correct and typically not considered plausible choices by examinees. d. The central idea of the question needed to be stated in the stem. e. Choices that were unequal in length (i.e., correct choice was either the longest or the shortest). f. Some questions were missing from the version sent on 9/3. 2. TASC p/d's response stated: I'm working on the Gr 8 Earth History and MicroLife tests to deliver to you tomorrow. I thought … (the tests authors) had sent you the % of time spent on the 16 Earth History and 8 Microlife instructional

Grade 8 MicroLife (ML) pilot test assembly and Grade 8 Earth History (EH) pilot test assembly*	
	objectives in the NCSCoS goals. Did we identify objectives being assessed for questions on the other kits? I'll do my best to get this to you tomorrow, but … (who wrote the Earth History test) is training tomorrow all day, and then we're both going to a conference … tomorrow night. [Test author for ML test] drew all of the Microlife questions from the manual. I'm not sure how well I'll do with fixing them, but I'll try. We do not want to use knowledge-level questions if we can avoid them. Tomorrow morning, if I can. For Earth History, I'll rewrite questions 2, 5, 7, 8 to be at least application and I'll rework questions to replace "all of the above" and "none of the above" responses, as in Earth History questions 2, 5, 7. I'll move Earth History concepts addressed to the stems, esp. Q 5 and 7. I'll also try to equalize the lengths of these questions. On the Earth History Test, I told you about questions 12-17 over the phone a few weeks ago. They actually are questions, but wierdly embedded in the diagram. I'll see what I can do to make them into questions that we can all recognize. For MicroLife, I'll see if I can rework questions 2, 3, 4, 6, 7, 8, 12,13 to get them above knowledge level. I'm doubtful on that. I can get rid of the MicroLife "all of the above" and "none of the above" response options in questions 2, 5, 8, and restate question 8 to put the question in the stem. My only question is, is there any of this that you could do better than me? Tomorrow is a full day and I want to get this back to you tomorrow. I know you need these tests in usable form so you can return them in time for us to get kit shipments out. So, to answer your questions about what I want to do regarding these tests: I'll do what I've said above, but I'd like you to pick whatever of the tasks I've outlined above you think you could do, let me know what they are, and help me with them. If you can't, just say so, and I'll do my best with the list above. 3. TB forwarded the TASC p/d's email to the CERE acting director.
9/15/05	1. TB emailed to TASC p/d that she had not yet received from the two TASC test writers the percentage of time spent on the 16 EH and 8 ML instructional objectives and that "[w]ithout this information, I cannot create a test blueprint; and without a test blueprint, I would not know what instructional objectives to address in writing items for these two tests." In this email, she also responded to the TASC p/d's second question, reminding him that she was "waiting for the printing shop to finish the

Grade 8 MicroLife (ML) pilot test assembly and Grade 8 Earth History (EH) pilot test assembly*	
	copying of the Gr 5 IWS and Gr 5 Ecosys tests—some time this afternoon. Then these tests need to be 'classroom packed'—something that takes 2-3 hours. In order to get them to [the TASC warehouse manager] by tomorrow (probably afternoon), they will need to be hand-carried [an hour trip one way]." 2. The CERE acting director responded to the TASC p/d's 9/14 email to TB: … I was a little disturbed by your latest on the earth history/micro life assessments. First of all, I realize that teachers were not trained to write the items for these components but item writing was not in Terry's original contract. She was to make sure that the items were psychometrically sound and to make sure that you receive the analysis of these items to assist in your evaluation. She truly has gone above and beyond the call of duty. She worked very closely with the teachers that stayed in the program long enough to submit items. Daily she would communicate with them and make recommendations. She has studied the science curriculum and made sure that these items are all aligned with the NCCOS. She made sure that all of your assessments were shipped and when you asked her to break them down by classroom, she made sure that they were counted by classroom. Matt, our graduate assistant, even drove them to your warehouse to make sure that they got shipped on time. You do realize that you are only paying Terry $10,000. for all of the work that she has done thus far on your project. We submitted a budget for her to work more hours but that budget was rejected. As you know, writing a *_good_* assessment items is very difficult and it is time-consuming work. To even ask her to write, edit and submit them to you in the time frame that you have requested seems a little absurd. I realize that you are under a time crunch and I understand that you need to have this information for your evaluation but to expect Terry to do all of this work and not compensate her for it does not seem fair. 3. The TASC p/d responded to the CERE acting director's email by stating that he "didn't expect it of her" and that he had "made it clear [he] was just asking if there was some of it she thought she could do more efficiently than me." 4. TB emailed the TASC p/d requesting verification of the instructional objective being assessed by the 14 EH questions and the 11 ML questions and indicating that "there do not appear to be enough questions on Goal 6 …, where 60% of the training occurred."

Grade 8 MicroLife (ML) pilot test assembly and Grade 8 Earth History (EH) pilot test assembly*	
	5. ML: TB emailed the TASC p/d with a link to the Glencoe Science textbook as a potential source of ML items. 6. ML: The TASC p/d emailed to TB that he would forward the information on to the TASC scientist who was the ML test developer. 7. ML: The TASC p/d emailed (4:51 pm) to TB that he was: bogged down on MicroLife, but here's my verification of your list of instructional objectives being assessed and NC thinking skill(s) used. I have to go to a conference … now and all day tomorrow, but I'll finish MicroLife first thing Monday. Hopefully, … [the TASC scientists] will have looked at that web site and have some good replacements. That MicroLife test, as is, is a real dog. Sorry I wasn't able to do more with it.
9/16/05	TB emailed to the TASC p/d the formatted version of the EH test, asking him to look at item 6 on which she had a question.
9/19/05	1. EH: The TASC p/d emailed to TB that he had talked to the EH workshop instructor about her question on item 6, stating that "the item is best left as is because the distractor … requires the test taker to think. …" 2. ML: The TASC p/d emailed the following to TB: Microlife is tomorrow, and I can't get to that test today. Therefore, it appears that we won't be able to administer the teacher pretest. I'm in Robeson Co. all day tomorrow. I can work on the MicroLife test Wednesday evening after I train, and get it to you then. Maybe we can work it out on Thursday, then get the tests ready to ship out in the kits the following week. 3. The TASC p/d emailed to the EH and ML workshop instructors, including TB in the "cc", the instructions to students that were to be printed out and included in the kits. He indicated that "[t]he teacher should instruct students to fill in the ID section of the test according to these directions." 4. EH: TB emailed to the TASC p/d asking if the EH test was ready to be prepared for printing. 5. EH: The TASC p/d responded that the EH test was ready for printing.
9/20/05	ML: Session I of two-day workshop (Fall Cycle). We missed the finalizing of the teacher test in time for it to be piloted on this date.
9/21/05	EH: Session I of two-day workshop (Fall Cycle): TASC made copies of the Gr 8 EH teacher pretest for the 9 workshop registrants.

Grade 8 MicroLife (ML) pilot test assembly and Grade 8 Earth History (EH) pilot test assembly*	
9/23/05	Gr 8 ML and Gr 8 EH curriculum units shipped out of TASC Resource Center.
9/26/05	ML: TASC p/d emailed to TB that he had revised the ML test and sent it to the TASC scientist teaching the workshop "to verify that the revisions match the kit and the standards in accord with the % time spent" and that he'd get back to her the next day.
9/27/05	1. The TASC p/d emailed to TB the revised (and confirmed) ML test along with the EH answer key. 2. TB responded to TASC p/d's email by acknowledging receipt of EH answer key and requesting verification of assessed instructional objective for the 15 questions on the ML test. She also requested revisions on items 2 and 4. Lastly, she indicated that "there (still) do not appear to be enough questions on Goal 6 …, where 60% of the training occurred. For a 15-question test, 9 questions should be on Goal 6, with 6 question s assessing 6.01 and 3 questions assessing 6.04. … we current have 5 questions assessing 6.01 [and] … still need 3 questions assessing 6.04 and at least one question assessing either 6.02 or 6.03."
10/3/05	ML: TASC p/d responded to TB's 9/27 email by sending a revised ML test and indicating the instructional objective assessed by each question as well as the NC thinking skill(s).
10/4/05	1. ML: TB emailed to the TASC p/d "the newly revised MicroLife test" and asking him to "take a looksee at the test and see if any additional revisions need to be made" and to "provide answers for any questions that do not have answers [underlined] …". 2. ML: The TASC p/d responded to TB that "[i]f there are any questions you could take out to better balance the test, that would probably be better, especially if you find any from goal 7 that look redundant, but it is OK to leave it as is." He also indicated that he had "underlined all of the correct answers, and on question 18 I reworded the correct answer …."

Grade 8 MicroLife (ML) pilot test assembly
and
Grade 8 Earth History (EH) pilot test assembly*

10/5/05

1. ML: TB responded to the TASC p/d's 10/4 email with the following table:

Instructional Objective	% of Time Spent on Objective	No. of Questions Needed if Total Items = 18	No. of Questions Needed if Total Items = 15	Items from ML test Covering Instructional Objective
6.01	30%	5	5	Q3; Q14
6.02	5%	0—1	0—1	Q7; Q12; Q13; Q16
6.03	5%	0—1	0—1	
6.04	20%	4	3	Q5; Q17; Q18
7.01	5%	0—1	0—1	
7.02	10%	2	2	Q2; Q4; Q15
7.03	20%	4	3	Q1; Q6; Q9; Q10; Q11
7.04	5%	0—1	0—1	Q8

She also stated:

I'm OK with having *more* items than we need to assess a particular instructional objective (e.g., 6.02, 7.02, 7.03) because we may find from pretesting that some of the items do not perform as expected and we may have to drop them.

However, based on this table, I think we still need *at least* 3 items for instructional objective 6.01—because 30% of the training time is being spent on this particular objective.

2. ML: TB sent to the TASC p/d a second email that replicated what he had sent to her in mid-September as to the percentage of instructional time spent on the instructional objectives.

10/6/05

ML: The TASC p/d responded that he and the ML kit instructor have arrived at a better estimate of MicroLife Kit training time as it is distributed across the NCSCoS goals. [The ML instructor] apologizes for the time we lost due to trying to accommodate innaccuracies [sic] in the previous estimate. I hope this helps. Let me know if this distribution means we'll need to write any more items. The MicroLife Kit training is next Tuesday.

10/7/05

ML: The TASC p/d emailed to TB that he thought the ML test needed no further changes and asking if there was something more he was supposed to do on it. He indicated that each goal appeared to match the percentage of time spent

Grade 8 MicroLife (ML) pilot test assembly and Grade 8 Earth History (EH) pilot test assembly*	
	training on it in the workshop and that "we just need for you to get it ready for printing, email it to us, and we'll photocopy it to give out as the pre-test for teachers." In his email, the TASC p/d also asked about the two versions of the test, stating: How will we go about making two separate tests, one for students and one for teachers? Will someone adapt the teacher-level test to student level? Who? Will we have student-levelled tests for the January cycle? Will we do it after we get data back on the items? I imagine that, especially for grade 3, the higher language level will cloud interpretation of the items' performance. How much of a problem is that?
10/10/05	1. ML: TB emailed to the TASC p/d the ML teacher pretest (sections 1 and 2), instructions for the teacher pretest, the teacher posttest, instructions for the teacher posttest, and the Scantron coding instructions. 2. ML: The TASC p/d acknowledged receipt of the 6 files and that he would make a pre-test package and a post-test package from the files. He stated: "We will have teachers take the pre-test tomorrow and then hand out the post-test for them to take when they finish the unit with their students. They will then place the completed post test in the kit when they return it."
10/11/05	ML: Session II of two-day workshop (Fall Cycle): TASC made copies of the Gr 8 ML teacher pretest for the 20 workshop registrants.
2/2-13/06	ML: In preparation for the 2/15/06 ML workshop, TB: • revised Pretest/Posttest Coding sheet; • revised instructions on student tests; • obtained 5,200 Scantrons and had 2,600 student pretests and 2,600 student posttests printed; and • assembled 26 "assessment packs" that included: o One “Pretest Directions to Students” to be read to the students, 100 student pretests with “General Pretest Directions to Students,” and 100 Scantron answer sheets. o One “Posttest Directions to Students” to be read to the students, 100 student posttests with “General Posttest Directions to Students,” and 100 Scantron answer sheets. o One “Posttest Directions to Teachers” to be read by the teacher before taking the posttest, and one Scantron answer sheet. o Two postage-paid, CERE-addressed envelopes: The teacher will use one

Grade 8 MicroLife (ML) pilot test assembly and Grade 8 Earth History (EH) pilot test assembly*	
	of these envelopes to return students' pretest Scantron answer sheets to CERE and one to return teacher and students' posttest Scantron answer sheets.
2/15/06	ML: Session I of two-day workshop (Winter Cycle): TB administered at the TASC Training Center the teacher pretest to workshop attendees and distributed an "assessment pack" to each attendee.
*These two tests are included in one table because initially were assembled simultaneously. Their assembly diverged when the ML session I deadline was missed.	

Table 30. Grade 3 *Human Body (HB)* pilot test assembly

Grade 3 Human Body pilot test assembly	
8/26/05	Using items written by IW02, TB created first draft of teacher and student tests.
9/13/05	TB sent drafts of tests to TASC p/d (i.e., 082605 Gr3 HB).
9/16/05	1. TASC t/d emailed to TB his version of the test: "I've created one version for both teachers and students." 2. TB acknowledged receipt of test and asked for a head count of teachers for the next week's workshops (i.e., Grade 3 HB, Grade 8 EH, Grade 8 ML). 3. TASC t/d responded with head counts. 4. TB emailed TASC t/d with two questions (clarifications) on item 2. 5. TASC t/d responded with revisions to HB test. 6. TB emailed revised HB test "with revisions incorporated and with choices arranged alphabetically. Please take a looksee and let me know if it's ready to be printed." 7. TASC t/d sent answer key to HB test. 8. TB sent email that she was going to take tests to printer that afternoon. 9. TASC t/d's response: "No reason to wait. I will not be in the office until late Monday, so if you have further questions, [TASC p/d] will have to try to answer them."
9/22/05	Session I of two-day workshop (Fall Cycle): TASC made copies of the Gr 3 HB teacher pretest for the 28 workshop registrants.

Grade 3 Human Body pilot test assembly	
9/23/05	Gr 3 HB curriculum units shipped out of TASC Resource Center.

Table 31. Grade 5 *Motion & Design (M&D)* pilot test assembly

Grade 5 Motion & Design (M&D) pilot test assembly	
8/31/05	Items were accepted by CERE from IW19.
11/4/05	TASC p/d emailed to TB the following: We did not write a Motion & Design test (5th grade) because we didn't train on it this quarter. However, Alamance County will offer a training on it in January, and we will offer it Feb. 14th. I'd like to get started on that one, and do it just as we did the others, if that's OK with you. Please let me know if we're ready to do that and how you'd like me to proceed.
11/22/05	TB emailed to TASC p/d drafts of student and teacher M&D pretests.
11/23/05	TASC p/d acknowledged receipt of the tests and indicated that he would look them over and get back to her about them the following week.
11/28/05	TASC p/d emailed to TB his changes to the teacher pretest.
11/30/05	1. TB emailed to the TASC p/d the second draft of the M&D test. She stated that she had incorporated all his revisions and that she had included items 11 through 23 for his review. 2. The TASC p/d emailed back to TB, indicating he had looked over the test and found a needed correction for item 12. He stated: "Otherwise, it looks good."
12/1/05	Per the TASC program evaluator's (p/e) request, TB emailed a copy of the teacher pretest, asking "will this test be piloted in early January 2006 with the 20 Robeson County teachers you were telling me about?"
12/2/05	The TASC p/e responded to TB's 12/1 email as follows: Yes, we would like to pilot test this test (Motion and Design) when the Robeson Co. teachers come back in January which means we need to send the tests out soon. . . . Also a group of teachers in Alamance County will receive training in January as well and I

Grade 5 Motion & Design (M&D) pilot test assembly	
	would like to field test it on them too. . . .
Week of 1/2/06	The TASC p/e administered the 23-item M&D teacher pretest to 23 Robeson County teachers.
2/3/06	TB met with the TASC p/d to discuss, among other things, the M&D test. In response to feedback from the TASC program evaluator's administration in January 2006 of the 23-item M&D test that it was "too long", the TASC p/d requested the test length not exceed 15 items. TB suggested the creation of two forms of the test, and he agreed.
2/7-13/06	In preparation for the 2/14 M&D workshop, TB: • revised Pretest/Posttest Coding sheet; • revised instructions on student tests; • created Forms A and B (with 11 anchor items) of M&D test (same version for both students and teachers); • obtained 1,080 Scantrons and have 540 student pretests and 540 student posttests printed; and • assembled 17 "assessment packs" that included: o One “Pretest Directions to Students” to be read to the students, 27 student pretests with “General Pretest Directions to Students,” and 27 Scantron answer sheets. o One “Posttest Directions to Students” to be read to the students, 27 student posttests with “General Posttest Directions to Students,” and 27 Scantron answer sheets. o One “Posttest Directions to Teachers” to be read by the teacher before taking the posttest, and one Scantron answer sheet. o Two postage-paid, CERE-addressed envelopes: The teacher will use one of these envelopes to return students’ pretest Scantron answer sheets to CERE and one to return teacher and students’ posttest Scantron answer sheets.
2/14/06	Session I of two-day workshop (Winter Cycle): TB administered at the TASC Training Center Forms A and B of the M&D teacher pretest to 17 workshop attendees and distributed an "assessment pack" to each attendee.

This researcher's feedback to TASC included:

- pointing out parts of an item that could be confusing to an examinee (e.g., The statement "The moon can be seen in the sky" was part of a stem. This researcher asked whether the correct response was dependent upon what *phase* the moon was in.);
- requesting information as to what instructional objective was being measured by a question;
- pointing out knowledge-level items;
- asking whether more than one alternative could be correct;
- pointing out very long (and correct) alternatives and/or very short alternatives; and
- providing readability levels and suggesting ways to break up compound sentences.

In addition to the assembling of the tests, test administration procedures were developed at the same time, e.g.:

- "ID Section Coding instructions" for teachers and students to use codes on the Scantrons so CERE would be able to match students to teachers to workshops;
- "Instructions to teachers" that TASC test administrators (i.e., TASC workshop scientists) were to read to the teachers prior to the pretest;
- "Instructions to students" that the teachers were to read to their students prior to the pretest and prior to the posttest;
- Survey questions for both students and teachers that were contained in Section Two of the pretests; and

- Color-coding tests to make them easier to separate upon their return; that is, white for teacher pretests, blue for teacher posttests, yellow for student pretests, and green for student posttests.

To summarize, the TASC project director and scientists reviewed, edited, and wrote items for the draft tests. The draft tests were then emailed to this researcher, who reviewed the TASC-revised tests and, using the same criteria she used to review the teacher item-writers' test questions, provided feedback. After TASC approval and this researcher's incorporation of the (final) revisions, she then checked the tests for grammatical and/or spelling errors, placed the choices into alphabetical order, and provided adequate "white space" throughout the test. The tests, then, were a compilation of items either written by the teacher item-writers and/or TASC scientists or obtained by this researcher from an approved source.

Once the item generation process was completed and the final pilot version emailed to the TASC project director, TASC made a copy for each workshop attendee. Meanwhile, this researcher, using a private copying company, had the student pretests, student posttests, and teacher posttests copied. The TASC project director and this researcher decided to color-code the tests to facilitate recognition of each test upon their return to TASC and then to CERE. That is, teacher pretests were copied onto white paper, student pretests onto yellow paper, teacher posttests onto blue paper, and student posttests onto green paper. At CERE, this researcher created "test packs" for each of the TASC workshop registrants. The packs included one large, sealed enveloped that contained student pre-tests (per TASC, 27 per teacher for Grades 3 and 5 and 100 per

teacher for Grade 8); one large, sealed envelope that contained student post-tests; and one sealed envelope that contained the teacher post-test. These "test packs" were delivered (either by delivery service or hand-carried by a CERE employee) to the TASC warehouse facility, where the TASC curriculum director and materials manager would add them to the science curriculum units before they were shipped out to the individual classrooms of the TASC workshop attendees.

Finally, whereas two versions—a student version and a teacher version—were to have been created, TASC dropped the student versions of the tests, again due to the severe time constraints. This meant that students and teachers were assessed using the same test version.

Factors that Affected Pilot Test Assembly

The predominant factors that affected the pilot test assembly task were the TASC project's participants, i.e., the teacher-item writers and the TASC scientists. Because the teacher-item writers struggled with the item writing, the time allotted for item writing was extended from three weeks to approximately eight weeks until a cut-off date (August 31, 2005) was set. This resulted in an extremely compressed timetable for CERE to assemble, print, and deliver the tests to TASC within its prescribed pilot testing deadlines.

In addition to taking more than twice the amount of time to write the items, the quality of the teacher-created items was a disappointment—and a surprise—to the TASC project director, who had thought the Training Institute teachers should have been "the

cream of the crop" as item-writers. In his interview with this researcher, who asked him "what surprised you about this [item-writing] process?", he stated:

> … how ill-equipped teachers are to write higher-order thinking skills items. In fact, how ill-equipped they are to ask questions even of … their students … that require them to think. … they're not used to setting up the game for the kid to where the kid has to struggle with the material. They're … so locked into the mode of getting the kid to give the right answer that … thinking … is an abstract thing. It's an abstraction to them, and it's something that … nobody … has worked with them, maybe with the exception of us, to try and develop … their ability to develop thinking in students. So I just think it's … something that's missing from teacher training, it's something that's … missing from their daily practice as teachers and all the sudden they were asked to do something that's basically outside their skill set.

Others, however, were not surprised. In his interview with this researcher, the TASC training director stated:

> I feel really strongly that the test developers need to know intimately the material that is being tested, and … relying on teachers to write the tests I would … *never* recommend … I'm not surprised that they wrote such poor questions and my assessment was that they wrote *terrible* questions for the most part … I'm not surprised by that knowing now what I know about their content knowledge.

Additionally, the TASC training director stated that "what I found was that those questions were *so* bad so often that the best I could do with them was still pretty awful, but that's what I did because of the limited amount of time I had."

Likewise, the TASC curriculum director, in his interview with this researcher, stated:

> … when it [item writing] initially started [the TASC project director] said that they were going to try to get teachers to do it. And I … knew that wouldn't work

> … because the teachers are really just getting started at using this kind of teaching and … it's really different and I think … for the broad range of teachers who come to the workshops it was beyond their capabilities to sort of see beyond how the kits really worked and what their teaching is like with these kits and how do you evaluate something like that? You know, it's just going to sort of be pretty cut and dried textbooky kind of test, not much thinking. And even then the questions wouldn't be very good. For a couple reasons. They don't understand the kits so well and they don't know very much science. So I think in part when they work with us they're still trying themselves to learn the science that's in that kit and to be able to see a question that's beyond, you know, what color did it turn when you added the drops, you have to know more.

When asked if he was surprised at teachers' lack of science knowledge, he responded, "not really", based upon his experience from working in the school systems.

The TASC materials manager, in his interview with this researcher, also expressed his opinion that teachers should not have been used as item writers because they were not

> killed in the area of writing test questions … in a way that it isn't just [to] regurgitate information but … in ways that there's some thinking going on, some process that has to be involved and the answers are such that you have to think about different answers and try to wrestle what is the answer that seems most appropriate. I'm not qualified for that and I don't think teachers are as well.

As stated at the beginning of this section, the extended time given to the teachers to complete their item writing task resulted in an extremely compressed timetable to assemble, check, revise, print, pack, and deliver the tests to TASC. In addition to teacher-created items that needed extensive revisions and/or replacement—all of which had to be done by TASC scientists, TASC's determination to include five tests for which no items had been written as of August 31, 2005 further complicated the test assembly

process. This, of course, translated into TASC scientists becoming fast-tracked item writers for tests that were to be piloted within two to three weeks. Even though the scientists clearly knew their content, they were inexperienced at writing grade-level multiple choice questions that matched test blueprints, were of medium to hard difficulty, and assessed higher order thinking skills. As already mentioned in the TASC materials manager's comment above, some scientists felt ill-equipped to write such items.

This extremely compressed timetable also resulted in dropping two test versions—teacher and student—and making one version for *both* teachers and students. While this may have been somewhat acceptable for the eighth grade tests, it seemed to be less appropriate for the fifth grade tests and inappropriate for the third grade tests.

Additionally, the extreme timetable resulted in missing the initial deadline (i.e., September 20, 2005) for the Grade 8 MicroLife test. While work continued on the test through the second half of September and beginning of October, the test was not finalized until October 10, in time for the second day of training on October 11. However, the curriculum unit shipment date had been missed and thus no student testing resulted from this initial teacher pretest.

Last of all, not only was the development/assembly of the tests rushed, but the printing of student pretests and posttests for each teacher of a workshop, the packaging of student pretests and student posttests—for each teacher—so that TASC would not have to do this, and the delivery of the packaged tests to TASC's warehouse so that the tests could be included in their science kits TASC would send out to the teachers who attended the workshops were also rushed.

Pilot Test Administration

The TASC's project director's August 30, 2005 email to this researcher, referred to in the previous subsection, best expresses the expectations for the administration of the pilot tests. As a reminder, the TASC program director responded to the CERE acting director with his understanding of the work to be completed under the (extended or new) TASC-CERE subcontract. The parts of that email pertinent to test administration are reproduced below:

> . . . The items will be tested in the first cycle this year. The tests must be ready to hand out on Sept. 20. To do that, TASC needs the tests a week before Sept. 20 to prepare them. Once we have pre- and post-results, CERE will analyze the tests to prepare them [to] use in evaluation. If it helps with your budget, TASC has some capability to put responses on Scantron sheets and program the scanning software. In that case, we could send you the sheets and the raw data for your analysis.
>
> TASC personnel will administer teacher pre-tests to teachers at TASC training. CERE/UNCG will receive pretest results the week pre-testing is completed. After the teachers complete using the unit with their students in about 9 weeks, they will self-administer the post-test. Kits will contain both pre-tests and post-tests for the students. Teachers will administer student pre-tests before they begin to use the curriculum units and administer the student post-tests after the science unit is completed about nine weeks later. . . .
>
> A summary of all the tests, by kit and grade level, to be provided from the sources listed above and analyzed by CERE is as follows:
>
> - Grade 3 Plant Growth & Development
> - Grade 3 Human Body
> - Grade 3 Soils (tentative)
> - Grade 4 Magnetism & Electricity (under discussion)
> - Grade 4 Food Chemistry (under discussion)
> - Grade 5 Landforms
> - Grade 5 Motion & Design
> - Grade 5 Ecosystems STC
> - Grade 5 Investigating Weather Systems (T.R.A.C.S.) BSCS

- Grade 8 Earth History FOSS-Delta (we already have this)
- Grade 8 Micro-Life SEPUP

From this email, we learn that TASC's expectation was that all 11 tests would be piloted in the first TASC training cycle of the 2005-2006 academic year (Figure 13, reproduced in the previous subsection). That is, TASC personnel would administer teacher pretests on Session 1 of the Fall Training Cycle and would send the test results to CERE "the week the pretesting is completed". After 9 weeks (or 14, depending on the unit), teachers would administer the student pretests and posttests and self-administer the teacher posttest and then place tests and answer sheets in the science kit to be returned to TASC. TASC, in turn, would retrieve the tests and answer sheets from the science units and have them delivered to CERE. CERE would then analyze the test results and report its findings to TASC.

Three sets of standardized test administration instructions were prepared by this researcher and provided to TASC for each test:

- instructions to teachers, to be read by the TASC workshop instructors prior to the teacher pretest;
- pretest instructions to students, to be read by classroom teachers prior to the student pretest; and
- posttest instructions to students, to be read by classroom teachers prior to the student posttest.

Unique identification codes were assigned to each teacher to enable this researcher to match students to their teachers. Instructions included the recording of these

identification codes as well as the recording of answers onto the answer sheets (i.e., Scantrons). Content questions were scored as correct or incorrect. In addition, located at the top of the first page of each test were the following test-taking directions:

- Carefully read each question and each of the answer choices.
- Decide which of the answer choices is most correct for that question.
- Then go to the Scantron for that question. Fill in the oval that matches the answer choice you selected.
- Be sure to answer every question. Your score on the science test will be all the questions you answer correctly.

The Grade 3 tests (Human Body, Investigating Objects in the Sky, and Plant Growth & Development), three of the four Grade 5 tests (Ecosystems, Investigating Weather Systems, and Landforms), and one of the two Grade 8 tests (Earth History) were piloted in September 2005. Teacher tests were administered by TASC personnel (i.e., workshop instructors) to the teacher-participants at the TASC Training Center prior to training on the particular curriculum unit. Student pretests—that had been assembled, printed, and delivered by CERE—were included in the TASC curriculum unit kits, along with the teacher and student posttests; these were to be administered by these teacher-participants in their respective classrooms.

In January 2006, the TASC project evaluator administered on her own the 23-item Grade 5 Motion & Design test to approximately 20 Robeson County teachers. Other than providing the test by email to the project evaluator, CERE was not involved in the administration or analysis of this test. However, teacher feedback from the evaluator's test administration was that the test was "too long!" In response, the TASC project

director requested this researcher to limit the length of the test to 15 items. In early February 2006, this researcher created two 15-item forms (with 11 anchor items) of the Grade 5 Motion & Design test.

The Grade 5 Motion & Design (Forms A and B) and the Grade 8 MicroLife teacher tests were piloted in February 2006. These tests were administered by the author to the teacher-participants at the TASC Training Center prior to training on the curriculum unit. This researcher prepared an “assessment pack” for each teacher-participant. These included the student pretests, teacher posttest, student posttests, administration instructions, prepaid envelopes and Scantron answer sheets, all placed in a TASC-provided canvas bag. In addition to written instructions to each teacher included in the "assessment pack," this researcher verbally instructed teacher-participants to administer the student pretests upon their return to their classrooms, to administer student posttests approximately 9 to 14 weeks later, and to self-administer the teacher posttest at the same time as the student posttest. She also verbally reminded the teachers that prepaid envelopes were provided in the "assessment packs" for teachers to return all Scantron answer sheets from the pretesting and posttesting and to return the science tests upon completion of the posttesting.

Table 32, the pilot testing schedule, documents the actual dates of the teacher pretests and the number of teachers who took the pretest. Estimated student numbers were calculated based on TASC's estimate of 27 students per (elementary school) teacher and 100 students per (middle school) teacher. The estimated length of time that units were in a participating teacher's classroom was set by TASC. The estimated date of the

student pretest was based on the approximate first week that TASC science units would have been in the teachers' classrooms. The estimated date of the posttest was based on the approximate week that TASC science units would have been returned to TASC.

Table 32. Pilot testing schedule

Science Test	No. of Items	Date of Teacher Pretest	Actual No. of Teachers	Estimated Number of Students	Expected Length of Time Unit in Class-room	Estimated Date of Student Pretest	Estimated Date of Posttests
Gr 3 Soils[1]	n/a	9/7/05--cancelled	0	n/a	n/a	n/a	n/a
Gr 3: Investigating Objects in the Sky	13	9/7/05	17	459	9 wks	Wk of 9/12/05	Wk of 11/14/05
Gr 3: Plant Growth & Development	13	9/13/05	9	243	14 wks	Wk of 9/19/05	Wk of 1/2/06
Gr 5: Landforms	15	9/13/05	18	486	9 wks	Wk of 9/19/05	Wk of 11/14/05
Gr 5: Investigating Weather Systems	17	9/14/05	8	216	9 wks	Wk of 9/19/05	Wk of 11/14/05
Gr 5: Ecosystems	10	9/15/05	11	297	9 wks	Wk of 9/19/05	Wk of 11/14/05
Gr 4 Magnetism & Electricity[2]	n/a	9/20/05	n/a	n/a	n/a	n/a	n/a
Gr 8 MicroLife[3]		9/20/05 expected; 10/11/05 --actual	20	0	14 wks	n/a	n/a
Gr 8: Earth History	14	9/21/05	9	243	14 wks	Wk of 9/26/06	Wk of 1/9/06
Gr 3: Human Body[4]	10	9/22/05	28	756	9 wks	Wk of 9/26/05	Wk of 11/28/05
Gr 5 Motion & Design[5]	23	Wk of 1/2/06	23	0	n/a	n/a	n/a
Gr 4 Food Chemistry[6]	n/a	2/8/06	n/a	n/a	n/a	n/a	n/a

Science Test	No. of Items	Date of Teacher Pretest	Actual No. of Teachers	Estimated Number of Students	Expected Length of Time Unit in Class-room	Estimated Date of Student Pretest	Estimated Date of Posttests
Gr 5 Motion & Design—Form A[7]	15	2/14/06	9	243	9 wks	Wk of 2/20/06	Wk of 4/24/06
Gr 5 Motion & Design—Form B[7]	15	2/14/06	8	216	9 wks	Wk of 2/20/06	Wk of 4/24/06
Gr 8 MicroLife[8]	18	2/15/06	19	1900	14 wks	Wk of 2/20/06	Wk of 5/29/06

[1]Soils workshop cancelled 9/6/05 due to low registration.
[2]Magnetism &Electricity test was dropped by TASC on 9/14/05.
[3]The 9/20/05 testing date was missed; TASC wanted the test piloted on the teachers at the Session 2 workshop. The consequence of this decision was that no student pretests/posttests were shipped with the curriculum units.
[4]TASC administered the 082605 version of the teacher pretest, rather than the 091605 final version. The 091605 version was used for the student pretests and the posttests that shipped with the curriculum units.
[5]This test administration took place independent from CERE; that is, the project evaluator conducted this administration.
[6]Food Chemistry test was dropped from the 2005-2006 TASC-CERE subcontract so even though items were developed, they were never used.
[7]TASC requested, based on feedback from the project evaluator's January 2006 administration, that the 23-item Motion &Design test be reduced to 15 items. Therefore, two forms, with 11 anchor items, were created for pilot testing purposes.
[8]Because the fall 2005 deadlines had been missed, TASC wanted a full piloting of this test during the winter cycle training.

A few observations can be made from Table 32. One observation is that although three tests were either cancelled (e.g., Soils) or dropped (Magnetism &Electricity and Food Chemistry), ten tests were piloted--seven tests in fall 2005 and three tests in winter 2006. Another observation is that, out of the ten tests, only one initial pretest deadline (for MicroLife) was missed. The consequence of missing the initial deadline was that no student tests were shipped with the curriculum units that went out in September 2005. Thus, even though TASC wanted teachers pretested at the second session of the fall 2005 workshop, there were never any student data associated with that administration.

However, a full piloting of the MicroLife test took place in February 2006, along with the piloting of the two forms of the Motion & Design (M&D) tests. A third observation is the mix-up with the Human Body test. As noted in the table, TASC administrators used the initial draft version of the test for pretesting teachers. The tests that were shipped were the final revised version of the HB test. This researcher discovered the error when she began scoring teachers' pretests and noticed too many teachers scoring items incorrectly. A final observation is the use of the initial 23-item Motion &Design test by the project evaluator in order to obtain "quantitative data" to include in her annual evaluation report.

Pilot data from the fall 2005 administrations of Grade 3 Human Body and Grade 5 Ecosystems arrived at CERE the week of December 19, 2005. Pilot data from the fall 2005 administrations of Grade 3 Investigating Objects in the Sky, Grade 3 Plant Growth & Development, Grade 5 Investigating Weather Systems, Grade 5 Landforms, and Grade 8 Earth History arrived at CERE the week of January 9, 2006. Student pretest data from the February 2006 administrations of Grade 5 Motion & Design (forms A and B) and Grade 8 MicroLife arrived at CERE in mid-March 2006, and student and teacher posttest data arrived at CERE throughout May and into June 2006.

Tables 33 through 43, arranged in the same order as the tests in Table 32, present each workshop attendee's actual classroom administration of the pilot tests. The "notes" column of each table documents the condition of the data received by CERE, particularly from the fall 2005 test administrations, and the variety of problems encountered, e.g., no pre- and/or post-test data received (which may have been due to teachers not using the

units in their classrooms); undated tests; tests with missing identification numbers making it impossible to match students to their teachers; different lengths of time—when it could be calculated—between pretests and posttests (from as short as 2.5 weeks up to 8.5 weeks); students not recording their responses on the Scantrons (which meant that Scantrons had to be filled out for these students).

Table 33. Grade 3 *Investigating Objects in the Sky* pilot tests (fall 2005)

IDNum	Date of Student PreTest (exp. wk of 9/12/05)	Date of PostTest (exp. wk of 11/14/05)	Notes	No. of Wks between Pre- and Post-tests (exp. 9 wks)
1*	9/19/2005	11/10/2005		7 wks
2			No data received	
3			No data received	
4*	10/3/2005	10/28/2005		3.5 wks
5*	9/19/2005	undated		??
6*		11/16/2005	No student data received	
7			No data received	
8	10/24/2005	11/17/2005		2.5 wks
9*	9/20/2005	11/1/2005		6 wks
10*	9/15/2005	11/7/2005		7 wks
11*	??	11/9/2005	No student pretest data at least that could be matched to this teacher; students did not fill in Scantrons on post-tests	
12			No data received	
13*	10/10/2005	11/16/2005		5 wks
14			No data received	
15*	9/15/2005	11/10/2005		8 wks
16			No data received	
17			No data received	
*(no ID)**	9/16/2005	11/16/2005	Teacher ID number unknown; assigned numbers 9999999078-100 to student posttests	8.5 wks
*(no ID)**			No student data received	
(no ID)	9/22/2005		Teacher ID number unknown; assigned numbers 9999999062-077 to student data	

IDNum	Date of Student PreTest (exp. wk of 9/12/05)	Date of PostTest (exp. wk of 11/14/05)	Notes	No. of Wks between Pre- and Post-tests (exp. 9 wks)
(no ID)	9/16/2005		Teacher ID # unknown; assigned numberss 0000000001-019 to student data	

Notes:
*teacher took posttest (n=11)

1. One set of student pre-test and post-test data with NO teacher IDs (so numbered them 9999 999 001 through 018).
2. Two sets of student pre-test data with no student names and no teacher IDs (so numbered them 9999 999 019 thru 039 and 9999 999 040 thru 061).
3. Four sets of student post-tests with completely blank Scantrons.
4. One set of student post-tests where the tests were filled in but not the Scantrons.

Table 34. Grade 3 *Plant Growth & Development* pilot tests (fall 2005)

IDNum	Date of Student PreTest (exp. wk of 9/19/05)	Date of PostTest (exp. wk of 1/2/06)	Notes	No. of Wks between Pre- and Post-tests (exp. 14 wks)
1*	none given	11/8/2005		
2			No data received	
3			No data received	
4*	9/21/2005	11/15/2005		8 wks
5*	9/21/2005	11/15/2005		8 wks
6			No data received	
7*			No student data received	
8			No data received	
9	9/22/2005	11/9/2005		
Unknown	9/22/2005	11/9/2005		7 wks
No ID#	9/27/2005	11/17/2005	Teacher ID unknown; assigned numbers 0000000001—0000000017 to student data	7 wks
No ID#	10/20/2005	12/13/2005	Teacher ID unknown; assigned numbers 0000000101—0000000118 to student data	7 wks

*teacher took posttest (n=4)

Table 35. Grade 5 *Landforms* pilot tests (fall 2005)

IDNum	Date of Student PreTest (exp. wk of 9/19/05)	Date of PostTest (exp. wk of 11/14/05)	Notes	No. of Wks between Pre- and Post-tests (exp. 9 wks)
1			No data received that could be matched to this teacher	
2*	undated	undated		
3*	undated	undated		
4*	9/22/2005	11/14/2005		7 wks
5	9/22/2005	11/14/2005		7 wks
6*	11/4/2005	11/16/2005		2 wks
7*	undated	undated		
8			No data received that could be matched to this teacher	
9			No data received that could be matched to this teacher	
10*	undated	11/3/2005		
11*			No student data received	
12			No data received that could be matched to this teacher	
13*	9/20/2005	11/17/2005		8 wks
14*	undated	undated		
15			No data received that could be matched to this teacher	
16			No data received that could be matched to this teacher	
17			No data received that could be matched to this teacher	
18*	11/4/2005	undated		

Note:
*teacher took posttest (n=10)
Received 3 sets of student pretests with no teacher ID numbers and one set with no student names.

Table 36. Grade 5 *Investigating Weather Systems* pilot tests (fall 2005)

IDNum	Date of Student PreTest (exp. wk of 9/19/05)	Date of PostTest (exp. wk of 11/14/05)	Notes	No. of Wks between Pre- and Post-tests (esp. 9 wks)
1*			No student data received	
2*	9/26/2005	11/4/2005	• Student pretests: 24 out of 43 students recorded answers on notebook paper; data had to be transferred to Scantrons.. • Student posttests: 24 out of 25 students recorded answers on notebook paper; data had to be transferred to Scantrons	6 wks
3			No data received	
4*	undated	undated		
5			No data received	
6	??	11/16/2005	• No student pretests could be matched to this teacher • Three sets of students' posttests filled in on a copy of the test; no Scantrons	
7	??	undated	No student pretests could be matched to this teacher	
8	undated	none given	• Students recorded answers on copies of Scantron sheet; all student data had to be transferred to original Scantrons • No teacher post-test; no student post-tests	
assigned 9998	undated	none given	Teacher not listed as IWS registrant; may have been a substitute for one of the teacher-registrants	
assigned 9999	undated	none given	Teacher not listed as IWS registrant; may have been a substitute for one of the teacher-registrants. Seven out of 16 Scantrons were on a copy of a Scantron so data had to be copied onto original Scantrons	

IDNum	Date of Student PreTest (exp. wk of 9/19/05)	Date of PostTest (exp. wk of 11/14/05)	Notes	No. of Wks between Pre- and Post-tests (esp. 9 wks)
Unknown*	undated	undated	No teacher ID number and she did not respond to emails. Assigned numbers 0000000100—0000000119 to student data	
*teacher took posttest (n=4)				

Table 37 . Grade 5 *Ecosystems* pilot tests (fall 2005)

IDNum	Date of Student PreTest (exp. wk of 9/19/05)	Date of PostTest (exp. wk of 11/14/05)	Notes	No. of Wks between Pre- and Post-tests (exp. 9 wks)
1*	10/19/2005	11/17/2005		4 wks
2	undated	undated		
3	undated	undated		
4	undated	11/16/2005		??
5	undated	undated		
6*	undated	undated		
7	undated	undated		
8*	undated	undated		
9*	undated	undated		
10	9/30/2005	11/15/2005		6 wks
11	undated	undated		
*teacher took posttest (n=4)				

Table 38. Grade 8 *Earth History* pilot tests (fall 2005)

IDNum	Date of Student PreTest (exp. wk of 9/26/05)	Date of PostTest (exp. wk of 1/9/06)	Notes	No. of Wks between Pre- and Post-tests (exp. 14 wks)
1			No data received	
2*	undated	undated		
3*	undated	undated		

IDNum	Date of Student PreTest (exp. wk of 9/26/05)	Date of PostTest (exp. wk of 1/9/06)	Notes	No. of Wks between Pre- and Post-tests (exp. 14 wks)
4*	undated	undated		
5			No data received	
6*	11/1/2005	12/20/2005	Posttest: Students in one section recorded part of their own SSN; all their numbers had to be erased and changed to reflect teacher ID number	7 wks
7	undated	none given		
8*	undated	undated		
9	undated	none given		
*teacher took post-test (n=5)				

Table 39. Grade 3 *Human Body* pilot tests (fall 2005)

ID Num	Date of Student PreTest (exp. wk of 9/26/05)	Date of PostTest (exp. wk of 11/28/05)	Notes	No. of Wks between Pre- and Post-tests (exp. 9 wks)
1*	10/12/2005	undated		
2	10/20/2005	none given		
3*	undated	undated		
4			No data received; teacher wrote on back of envelope: "I already did one of these during training."	
5*	10/10/2005	11/21/2005		6 wks
6	10/27/2005	none given		
7			No data received	
8			No data received	
9	10/5/2005	none given		
10			No data received	
11*	NONE given	undated	No student data received	
12	10/13/2005	11/16/2005		5 wks
13*	10/10/2005	11/17/2005		6.5 wks
14*	10/14/2005	11/17/2005		5 wks
15*	11/1/2005	11/21/2005	No student data received	3 wks
16*	10/10/2005	10/26/2005		2.5 wks
17			No data received	
18	10/10/2005	none given	No ID number on student	

ID Num	Date of Student PreTest (exp. wk of 9/26/05)	Date of PostTest (exp. wk of 11/28/05)	Notes	No. of Wks between Pre- and Post-tests (exp. 9 wks)
			pretests; assumed PW (written on yellow test forms) was LW	
19	10/11/2005	none given		
20*	10/10/2005	10/27/2005		2.5 wks
21*	10/10/2005	10/26/2005		2.5 wks
22*	10/11/2005	11/21/2005		6 wks
23*	undated	undated		
24*	10/6/2005	10/28/2005		3 wks
25*	10/17/2005	11/7/2005		3 wks
26*	undated	undated		
27*	10/17/2005	11/10/2005		3.5 wks
28*	10/10/2005	11/21/2005		6 wks
??	undated		No ID number on student pretests and nothing was written on yellow pretests so I was unable to match student data with teacher.	

*teacher took posttest (n=17)

Table 40. Grade 5 *Motion & Design-Form A* pilot tests (winter 2006)

ID Num	Date of Student Pretests	Date of Posttests	Notes	No. of Wks between Pre- and Post-tests
1			No student data received	
3*	2/22/2006	4/27/2006		9
5*	2/23/2006	4/24/2006		8
7*	2/15/2006	4/26/2006		10
9*	2/24/2006	5/4/2006	Pretests: entire 10-digit ID# field was filled in--had to erase last 3 digits in order to assign a unique number to each student; testdate was missing--I assigned 2/24/06.	10
11*	2/24/2006	4/13/2006	Pretests: teacher ID number not included on students' Scantrons; also testdate was missing--I assigned 2/24/06 as testdate (per memo to teacher)	7
13	2/21/2006	4/11/2006		7

ID Num	Date of Student Pretests	Date of Posttests	Notes	No. of Wks between Pre- and Post-tests
15	2/17/2006		No posttests received	
17		4/24/2006	No student pretests received	
*teacher took posttest (n=5)				

Table 41. Grade 5 *Motion & Design-Form B* pilot tests (winter 2006)

ID Num	Date of Student Pretests	Date of Student Posttests	Notes	No. of Wks between Pre- and Post-tests
2	2/21/2006	5/30/2006		14
4*		4/24/2006	No student pretests received	
6*	2/22/2006	4/27/2006		9
8	3/1/2006	4/27/2006	Posttest: 7 did not bubble in ID number and Special Code field; 36 out of 110 were filled in on a copy of Scantron--had to be recopied onto original Scantrons	8
10*	2/15/2006	3/17/2006		4
12*		4/24/2006	No student pretests received	
14	2/16/2006	4/3/2006		6
16*	2/24/2006	4/27/2006		10
*teacher took posttest (n=5)				

Table 42. Grade 8 *MicroLife* pilot tests (winter 2006)

ID Num	Date of Student Pretests	Date of Student Posttests	Notes	No. of Wks between Pre- and Post-tests
1	2/21/2006	5/18/2006		12 wks
2	2/22/2006		No posttest data	
3	2/24/2006		No posttest data	
4*	2/20/2006	4/24/2006	Pretest: had to correct ID numbers recorded as 1217 to 1712 on 27 Scantrons	9 wks
5	2/20/2006		No posttest data	
6*	2/21,22/2006	05/__/06		8 wks ("Apr 06--May 06")

ID Num	Date of Student Pretests	Date of Student Posttests	Notes	No. of Wks between Pre- and Post-tests
7*	2/20/2006	4/28/2006		9.5 wks
8*	2/28/2006	5/22/2006		11 wks
9*	02/23,24/06	5/22/2006		12 wks
10	2/20, 24/06		No posttest data	
11	2/17/2006		• Received pretest data 4/3/06 • No posttest data	
12	2/20/2006	4/10/2006		7 wks
13*	2/21/2006	4/27/2006		9 wks
14*	3/7/2006		No posttest data	
15*	3/15/2006	4/28/2006	Received pretest data 3/29/06	6 wks
16	2/23/2006	4/28/2006		9 wks
17*	4/18/2006	5/23/2006	Received pretest data 4/25/06	4 wks
18	2/22/2006		No posttest data	
19*	2/17,21,24,06	5/1/2006		9-10 wks

*teacher took posttest (n=10)

Table 43 provides examinee counts of teachers and of students who took the pretests and posttests. In addition, it also provides the counts of students for whom this researcher was able to match pre- and posttests. A few observations from Table 43 are worth noting: The first is the difference in the estimated number of students (based on Actual No. of Teachers x 27 (for grades 3 and 5) or x 100 (for grade 8)) and the actual number of students who received the pretests. For instance, in the case of the Grade 3 Plant Growth and Development, the number of students who received the pilot pretest was about one-third of the estimated number of students. Next, it should be noted that approximately half the teachers who took the pretest did not take the posttests. Teachers wrote notes such as: "I already took this test!" Most, however, would simply return the

teacher posttest in its unopened envelope. Lastly, you will note the number of students for whom this researcher had complete records—i.e., pretests *and* posttests. One of the worst pre-post matches was Grade 5 Investigating Weather Systems with less than 50 percent of student pretests matched to posttests. This may have been the test where the TASC test administrator told teachers that student names were not needed on the tests.

Table 43. Number of examinees for pilot testing

Science Test	Actual Date of Teacher Pretest	Actual No. of Teachers Pretested	Actual Number of Teachers Posttested	Actual Number of Students Pretested	Actual Number of Students Posttested	Number of Students with Matching Pre- and Post-tests
Gr 3: Investigating Objects in the Sky	9/7/2005	17	11	294	226	**191**
Gr 3: Plant Growth & Development	9/13/2005	9	4	83	101	**73**
Gr 5: Landforms	9/13/2005	18	10	341	287	**263**
Gr 5: Investigating Weather Systems	9/14/2005	8	4	133	174	**63**
Gr 5: Ecosystems	9/15/2005	11	4	244	221	**199**
Gr 8 MicroLife	10/11/05	20	9	n/a	n/a	n/a
Gr 8: Earth History	9/21/2005	9	5	352	222	**202**
Gr 3: Human Body	9/22/2005	28	12	425	286	**262**

Science Test	Actual Date of Teacher Pretest	Actual No. of Teachers Pretested	Actual Number of Teachers Posttested	Actual Number of Students Pretested	Actual Number of Students Posttested	Number of Students with Matching Pre- and Post-tests
Gr 5 Motion & Design—Form A	2/14/2006	9	5	156	156	**132**
Gr 5 Motion & Design—Form B	2/14/2006	8	5	242	292	**211**
Gr 8 MicroLife[8]	2/15/2006	19	11	1601	1056	**1027**

This researcher received little direct feedback about the test administrations. However, the TASC project director, after the curriculum director had administered the Grade 3 Plant Growth & Development teacher pretest, emailed to this researcher the following:

> [The curriculum director] told me in his 2nd Plant Growth & Development training session, 4 separate teachers, in separate instances, approached him about the fact that the Plant Growth & Development pre-test was too difficult for 3rd graders. They were careful to say that it was not the subject matter about plant growth that threw the students, but rather the reading level of the questions.
>
> I just wanted to pass that on and to see if t here is some way we can lower that reading level in the next iteration.

This researcher had attempted to make this point with TASC at the time of the test's development. However, in the interest of time, the student version was dropped (along with all the other student versions).

Factors that Affected Pilot Test Administration

Pilot test administration was most affected by the TASC's project participants—predominantly, the TASC-participating teachers and the TASC scientists. The teachers appeared to be reluctant participants in the pilot test administrations. One reason for this may have been the lack of specific information about their role in the project's evaluation. In this researcher's interview of the TASC training director, he stated teachers were simply told "we are going to ask for you to do pre- and post-tests with your students and we're going to give you some pre- and post-tests … when we get them developed. That's probably about all we ever said to them." When asked whether any mention had been made that the reason for the tests was to help evaluate the program, he responded:

> the reason … was two-fold—one was because NSF requires us to do that. They were a research organization and … this is part of their research … to test the effectiveness of programs like ours and the other thing we told them was and this is going to help us make a better program. We want to know how we're doing so formative assessment in that sense. … We want this to help us to understand what we need to do to change.

When asked whether the teachers were told that test results would not be used for accountability purposes, he responded,

> Yes, yes, yes, yes, yes, we were absolutely clear about that. … That we never took their names, that we would ask them like for example on our surveys and things that we handed out we would ask for the last four digits of their Social Security numbers but no more because we just wanted to be able to track this test and this test and this test all came from the same person but we don't want to know who you are. Absolutely. And that none of this would be reported to administrators or you know any of that stuff, absolutely not. This was … made very clear … at least I thought we were.

This researcher, in her interview of the TASC project director, also asked him "how much information was provided to teachers regarding their part in the project evaluation?" He responded "just … the material that you passed out [at the item writing workshop] and … I think I spoke to them one time." This response reflects the teachers' item-writing role, rather than their role in the project evaluation per se.

Some of the TASC scientists also appeared to be reluctant participants in the testing process. When asked by this researcher about the level of support he received from his staff concerning the tests, the TASC project director stated:

> [They were] always questioning why are we doing this, this seems so stupid, … you know … nobody's going to care about this. … The only thing that principals and superintendents … care about is the EOGs; they're not going to care about some internal test we've developed.
>
> My answer to them was always I care about this because I want to know if we're doing our job internally and I want to be able to revise our training, you know, … to do a better job. Also … if teachers aren't learning content that we say we're teaching, that's the first thing to know. … and … I think the guys are thinking, look, forget about that; all we care about at this point is that they even use the darn kits, and in a way, they're right. Because … what I tend to forget is … the hellishness … of the school environment that they're [the teachers] are going back into when they leave training. … in fact if they get to use the kit at all or if they do find a way to do it, even though we've provided everything, … that we're … ahead of the game. So … almost all the time they're only doing part of the kit … if they do use it. And something like a third of the time in the first years they don't use the kit at all. So, they're thinking why are we worried about whether we're teaching them the content, and I have to say, I think we're worried about whether we're teaching them the content because that's one key component [for them to be] comfortable using the kit and it's another key component in making sure that the students actually learn the material. So, … yes, my staff was reluctant. …They … knew … it was written into the proposal from the beginning, … but I was the one that actually had [the] desire.

When asked about his discussions with his staff about the tests and the testing process, the TASC project director responded:

> Yeah, yeah, that was the testing process mostly, what a minute, how much time are we going to have to spend on this? You know, at the end of this training? Well, I don't have that kind of time. … we want to go really easy on the teachers so we want to let them out as soon as we can.

To determine whether this reluctance on the part of some of the scientists may have been reflected in *how* the tests were administered, this researcher asked one of the scientist (the TASC materials manager) how he introduced the test to his workshop attendees. He responded:

> When they came in, I told them we were gonna take a pretest and I read the letter talking about the program [the instructions to teachers] and I gave them the test and told them that I wasn't going to be evaluating them and this wasn't to see what they knew about science and I wasn't concerned about any individual's performance; it was more for me as an instructor to gauge how well I'm doing in terms of getting the content taught to the teachers. And then I explained for the posttest that they were to do it on their own and ship it back with the student tests.

When asked whether the teachers made any comments to him after taking the test, he responded: "Nope. We just went right into the workshop."

This researcher also asked the TASC curriculum director how he introduced the tests to his workshop attendees. He responded:

> I read the intro material in the cover and I explained to them in addition in my own words that these were things where we … really wanted to know where they were coming from before they started this and please don't worry if you don't know things … our goal here is to cover this material in the next two days … of the workshop, or one day, and … that then we wanted them to take this same test

> again when they were done using the unit at the end of the 7 or 8 weeks and that we wanted them to give this same test, pre- and post-, to their students. And … I said … this is really important for us to understand how we're doing and to try to figure out what we need to do to get better.

In response to the question about whether he supported the test development process, he responded:

> Well, this is … sticky for me. … I did not understand the role that, I did not anticipate the role that I ended up having to play in this process. … that is, I thought that somebody else was developing the questions and that … my role was going to be just to sort of take a quick look and say, well, did this cover you know what we did. … I, in the beginning, I assumed … though I don't suppose I was ever told this I just sort of assumed that the test items would be developed by your Center. … I, in the beginning, did not understand that an attempt was going to be made to have the teachers write the questions, for example—I didn't know that; I learned that later.

In addition, this researcher asked the TASC project director how he introduced the test to his workshop attendees. He responded that he told them:

> This is not to evaluate you, your name won't be associated with this, … we need an identifier on here so put your last four digits of your Social Security Number but that's only to match the pre and post and … to match you with your students when you give your students this test. I said … don't worry about how you do on this … what we're trying to find out [is] what you're learning from us so that we can do a better job of teaching you …

In response to the question as to whether he read verbatim the test administration directions, he stated "I did. I did read those out loud; I actually read them."

Thus, although some of the TASC scientists may have been reluctant participants in this testing process, when asked about their test administrations, they indicated

following the established test administration protocol. Even so, the condition of the data received by CERE from the classroom administrations of the fall 2005 tests was poor.

For instance, in a January 6, 2006 email to the TASC project director regarding the Grade 3 Human Body test data, this researcher wrote:

> Out of 28 teachers who attended the TASC training for the Grade 3 Human Body, I received:
>
> - complete data (defined as teacher post-test with students pretest and students posttest data) from 14 teachers;
> - no data from 4 teachers;
> - one set of unidentifiable student pretest data (that is, there was no teacher ID number on any of the Scantrons and nothing was written on the yellow pretests to identify the teacher);
> - student pretest data only (i.e., no posttest data) from 5 teachers; and
> - incomplete posttest data (either teacher only or students only) from 4 teachers, one of whom I had no student pretest data either; and
> - a Teacher Post-Test envelope with "I already did one of these during training" written on the back of it (while I was able to identify the teacher I had no matching student pretest data.
>
> All in all, the Scantrons were a real mess. In a few instances, teachers did not include the last 4 digits of their SSN on their posttest (if they took the posttest) so that had to be tracked down. Quite often, students bubbled their names vertically (using one column) or by randomly selecting columns. In many instances, the teacher's ID (last 4 digits of their SSN) was NOT included in the Identification # section of the Scantron. To obtain the number, I had to go through the tests themselves in the hope that at least one student would have identified their teacher.
>
> Lastly, you will note on the Excel sheet that the science units were not used for the same amount of time. The number of weeks between pretesting and posttesting varied from 2.5 weeks to 6 weeks.
>
> Presently, the students' Scantrons are being cleaned up (that is, name fields correctly bubbled in, ID number field bubbled in with unique student ID number I assigned to each student, Special Code field bubbled in with pretest and posttest dates (where available)) in preparation for their scanning by UNCG's Teaching and Learning Center.

Hopefully, the Grade 5 and Grade 8 data will be more complete.

The TASC project director responded in a January 9, 2006 email to this researcher, "Regarding your email about the sorry shape of the test data you're getting, I'd like to talk with you about ways we can improve on that from our end."

In a January 26, 2006 email to the TASC project director regarding the Grade 5 Investigating Weather Systems test data, this researcher wrote:

> Of all the boxes TASC sent, only one box was marked "Investigating Weather Systems". In it, there was:
>
> - one set of student post-tests with no teacher name and no teacher ID number;
> - one set of student post-tests with teacher ID number but no student names and no student pre-tests;
> - one set of student post-tests with answers recorded on the tests and not on the Scantrons--and thus no teacher ID number (teacher=Jan Brown);
> - one set of student pre-tests with answers recorded on duplicated copies of a Scantron (teacher= Mr. Meirring);
> - one set of student pre-tests with answers recorded on duplicated copies of a Scantron (teacher= Mrs. Scarboro who was not listed on the 9/14/05 workshop attendance roster);
> - one set of student pre-tests and post-tests with responses recorded on notebook paper rather than on the Scantrons; and
> - one set of student pre-tests from a Mrs. Kuess, also not listed on the 9/14/05 workshop attendance roster.

The TASC project director responded:

> OK. This just gets worse and worse. Tomorrow, when I get in and have my calendars, I'll call you to set up a meeting to see what we can do to change this situation for the next round.

This researcher and the TASC project director believed that the poor quality of the data being returned was a reflection of the teachers' reluctance to participate in the testing process. On February 3, 2006, this researcher met with the TASC project director at the TASC Training Center to discuss what we could do differently for the quickly approaching February 14 and 15 administrations of the Grade 5 Motion & Design tests and the Grade 8 MicroLife test. The following is this researcher's note to file:

02/03/06 Meeting with … (TASC Project Director)

- Suggested changes to pretest and posttest directions
- Suggested Registration attendance sign-in sheet with a random number assigned to each attendee
- As there are only two more pilots—Gr 5 Motion & Design and Gr 8 Microlife—I would administer the Teacher Pretest at the beginning of TASC's Session 1. That way I would be able to explain who I am; what the test is for; and why we need their cooperation and participation.
- Suggested ways to facilitate teachers' classroom testing process for Grade 5 Motion & Design pilot:
    - "Assessment Pack" would be handed to each TASC Workshop Registrant at the end of Session 1. Each pack would include:
        - One "Pretest Directions to Students" to be read to the students, 27 student pretests with "General Pretest Directions to Students," and 27 Scantron answer sheets.
        - One "Posttest Directions to Students" to be read to the students, 27 student posttests with "General Posttest Directions to Students," and 27 Scantron answer sheets.
        - One "Posttest Directions to Teachers" to be read by the teacher before taking the posttest, and one Scantron answer sheet.
        - Two postage-paid, CERE-addressed envelopes: The teacher will use one of these envelopes to return students' pretest Scantron answer sheets to CERE and one to return teacher and students' posttest Scantron answer sheets.
- Suggested ways to facilitate teachers' classroom testing process for Grade 8 Microlife pilot:
    - "Assessment Pack" would be handed to each TASC Workshop Registrant at the end of Session 1. Each pack would include:

        - One “Pretest Directions to Students” to be read to the students, 100 student pretests with “General Pretest Directions to Students,” and 100 Scantron answer sheets.
        - One “Posttest Directions to Students” to be read to the students, 100 student posttests with “General Posttest Directions to Students,” and 100 Scantron answer sheets.
        - One “Posttest Directions to Teachers” to be read by the teacher before taking the posttest, and one Scantron answer sheet.
        - Two postage-paid, CERE-addressed envelopes: The teacher will use one of these envelopes to return students’ pretest Scantron answer sheets to CERE and one to return teacher and students’ posttest Scantron answer sheets.
    - Review Gr 5 Motion and Design Pretest
    - Review Gr 8 Microlife Pretest
    - CEU credit: Is this contingent upon teacher attending the 2-day workshop only? If so, what is teacher’s incentive for using the science kit? Can CEU credit be tied to workshop attendance AND using the science kit?

Follow-up after meeting with [TASC project director]:

[TASC p/d] agreed to allow me to "give it a try" re: my administering the teacher pretests. I had explained that (1) there are only two more tests that need piloting—Gr 5 Motion & Design and Gr 8 MicroLife—and that we could at least try it; and (2) since I am an "outsider" and not actively involved in the teacher training and since I'm the only who will be using the data from the test, I may be less threatening to the teachers.

- I will administer the Gr 5 Motion & Design teacher pretest on 2/14/06. I will also distribute, and explain, the Motion & Design "assessment pack".
- I will administer the Gr 8 MicroLife teacher pretest on 2/15/06. I will also distribute, and explain, the MicroLife "assessment pack".
- [TASC p/d] suggested, and I agreed, that we would print ONE test since the pretest and posttest are the same. In addition, I will have the cover page of the test on UNCG letterhead.

. . .

Things that I MUST do this week:

Gr 5 Motion & Design

1. The Pretest/Posttest Coding sheet needs to be redone: Need instructions at the top: "Record your ID # and your LEA number. These numbers will be needed when you administer tests to your students and for your pretest and posttest."

2. I need to separate Section One and Section Two. Section One = test. Section Two = survey. Survey should be returned only with the Scantrons for the Pretest.
3. The Name, etc., lines need to be removed on page 1 of the test.
4. I need to redo instructions to reflect that students are NOT to write on the tests because they will be reused. I need to distinguish the Test (section one) from the Survey (section two). Instructions need to be on UNCG letterhead.
5. I need to create TWO forms—Form A and Form B—capping each form at 15 items.
6. I need to create assessment packs: 54 Scantrons + 27 Tests; one Teacher Posttest.
7. I need to purchase 50 prepaid Priority One envelopes. Verify that 27 Surveys and 27 Scantrons will not exceed the weight limit designated by USPS.
8. I need to have the Motion & Design test copied: 20 classrooms & 27 students/classroom = 540 tests and 540 surveys.
9. I need to have 1080 Scantrons, assembled in 20 "lots" of 54 Scantrons.
10. If teachers are to record their answers ON their Pretest and Posttest, then need to make the instructions clear.
11. I need to assemble the assessment packs: Large outer plastic envelope with labeled envelopes inside (Gr 5 Motion & Design Test; Gr 5 Survey; Teacher Posttest) with Memo to Teachers (which may also need to be redone to reflect the above changes).

<u>Gr 8 MicroLife</u>

1. The Pretest/Posttest Coding sheet needs to be redone: Need instructions at the top: "Record your ID # and your LEA number. These numbers will be needed when you administer tests to your students and for your pretest and posttest."
2. I need to separate Section One and Section Two. Section One = test. Section Two = survey. Survey should be returned only with the Scantrons for the Pretest.
3. The Name, etc., lines need to be removed on page 1 of the test.
4. I need to redo instructions to reflect that students are NOT to write on the tests because they will be reused. I need to distinguish the Test (section one) from the Survey (section two). Instructions need to be on UNCG letterhead.
5. I need to create assessment packs: 200 Scantrons + 100 Tests; one Teacher Posttest.
6. I need to purchase 54 prepaid Priority One boxes. Verify that 100 Surveys and 100 Scantrons will not exceed the weight limit designated by USPS.
7. I need to have the Microlife test copied: 26 classrooms & 100 students/classroom = 2600 tests and 2600 surveys.
8. I need to have assembled 5400 Scantrons assembled in 26 "lots" of 200 each.
9. If teachers are to record their answers ON their Pretest and Posttest, then need to make the instructions clear.
10. I need to assemble the assessment packs: Large outer plastic envelope—may require two of these--with labeled envelopes inside (Gr MicroLife Test; Gr 8

Survey; Teacher Posttest) with Memo to Teachers (which may also need to be redone to reflect the above changes).

One outcome from the changes to test administration procedures was a higher percentage of matched pre-post student data to estimated number of student examinees for the February 2006 test administrations (58 percent) compared to the fall 2005 test administrations (46 percent). This indicated more students participated and average percentage of matched pre-post student data was higher—79 percent for the February 2006 administrations compared to 68 percent for the fall 2005 administrations. Additionally, although student test data received from the February 2006 test administrations required cleaning, the data required less time to clean than the data from the fall 2005 administrations.

In summary, even though some of the TASC scientists reluctantly participated in the test administrations, it appeared that they did appropriately follow the established test administration protocol. Even so, teachers remained reluctant participants and this was evidenced by the poor quality of data received from the classroom administrations of the fall 2005 tests. TASC and CERE worked together to address teachers' reluctance, and some improvement—in the quality of the test data and in the greater numbers of students for whom matching pre-post data were received—resulted.

The next section presents the revisions to the pilot tests.

Pilot Test Revision

As indicated in the above section, the first test revision—even before the test was officially piloted—occurred with the Grade 5 Motion & Design test. As stated

previously, a 23-item test had been assembled by November 30, 2005. This version was used by the project evaluator on Robeson County teachers who said the test was too long. In response to this feedback, the TASC project director asked that the test length be limited to 15 items. This researcher, in early February 2006, created two 15-item forms, with 11 anchor items, of the Motion & Design test, that she administered at the February 14, 2006 Motion & Design workshop at the TASC Training Center in Durham, NC.

On June 26, 2006, this researcher forwarded CERE's summary report for TASC to include in their annual report to NSF. However, on July 17, 2006, the TASC project director emailed to this researcher that TASC had not yet sent their annual report to NSF and that he would like to include CERE's final evaluation report, if it was ready. In addition, he requested a meeting "to talk over whether we've got tests complete enough and strong enough to administer to teachers and students coming up in September." Specifically, the project director asked:

1) If you don't think the tests are ready, and more work on them is possible, are you willing or able to do any of it?
2) If more work on the tests is possible, what will it take for us to figure out what more needs doing?
3) Once we figure out what needs doing, if you're willing to do it, what are your time constraints?
4) Once we figure out what needs doing, if you're willing to do it, how much might it cost, and can we budget it?

On July 21, 2006, this researcher emailed a final draft of the TASC report to the CERE acting director for her perusal, comments, questions, etc. She indicated her

approval and CERE's final report was forwarded, by mail, to the TASC project director. As part of the final report, this researcher included “action items” TASC needed to address for each test. These actions items, included in Table 44, below, were based on results from the item analyses.

Table 44. Action items for TASC from CERE final report

Test	Action(s) to be taken by TASC
Gr 3 Human Body	o Items 1 and 6 should be examined and reworked or replaced. o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised? o Do more items need to be added to the test?
Gr 3 Investigating Objects in the Sky	o Items 6, 7, 8, 9, and 13 should be examined and reworked. o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised?
Gr 3 Plant Growth & Development	o Items 1, 3, 5, 6, 7, 10, and 12 should be examined and reworked. o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised?
Gr 5 Ecosystems	o Item 5 should be examined and reworked (or replaced). o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised? o Do more items need to be added to the test?
Gr 5 Investigating Weather Systems	o Items 1 through 8 and 10 through 14 should be examined and reworked (or replaced). o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised?
Gr 5 Landforms	o Items 2, 4, 6, 7, 9, 10, 11 and 14 should be examined and reworked (or replaced). o Does the test match its blueprint?

Test	Action(s) to be taken by TASC
	o Does the percentage of time spent on each objective need to be revised?
Gr 5 Motion & Design—Form A	o Items 1 through 7 and 9 through 13 should be examined and reworked (or replaced). o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised?
Gr 5 Motion & Design—Form B	o Items 1, 3, and 6 through 13 should be examined and reworked (or replaced). o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised?
Gr 8 Earth History	o Items 1, 6, 7, 8, 10, 11, 13, and 14 should be examined and reworked (or replaced). o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised?
Gr 8 MicroLife	o Items 2 through 9, 11, 12, 14, 17, and 18 should be examined and reworked (or replaced). o Does the test match its blueprint? o Does the percentage of time spent on each objective need to be revised?

On August 18, 2006, the TASC project director and this researcher met to discuss the action items listed in CERE's final report to TASC. On August 25, 2006, this researcher emailed the CERE acting director, summarizing the August 18 meeting at which the TASC project director stated the tests to be revised and administered for the 2006-2007 year would be the three third grade tests (i.e., Human Body, Investigating Objects in the Sky, and Plant Growth & Development) and four fifth grade tests (i.e., Ecosystems, Investigating Weather Systems, Landforms, and Motion & Design) created

by CERE. The two eighth grade tests (i.e., Earth History and MicroLife) would not be revised.

Figure 14, TASC's 2006-2007 training calendar, indicates the dates by which TASC would need to have the final revisions of these tests. Thus, the tests with immediate priority for revisions included Grade 5 (Investigating) Weather Systems (9/12/06), Grade 5 Landforms (9/19/06), Grade 3 Human Body (9/20/06), and Grade 5 Ecosystems (9/21/06).

2006-2007 TASC CURRICULUM UNITS		Fall Cycle Training Dates		Winter Cycle Training Dates	
Pre-K	"I Notice, I Wonder" Inquiry*	10/05/06			
K	Wood & Paper	09/06/06	09/27/06	02/07/07	02/28/07
K	Comparing and Measuring*	09/14/06			
K	Ant Homes Under the Ground*			02/08/07	
1st	Solids & Liquids	09/07/06	09/28/06		
1st	Balance & Motion*	09/07/06		02/13/07	
1st	Organisms	09/14/06	10/05/06	02/01/07	02/22/07
1st	Pebbles, Sand & Silt	09/13/06	10/04/06	01/30/07	02/20/07
2nd	Changes	09/21/06	10/12/06		
2nd	Lifecycle of Butterflies*	09/14/06			
2nd	Sound	09/13/06	10/04/06	02/07/07	02/28/07
2nd	Air & Weather	09/06/06	09/27/06	02/06/07	02/27/07
3rd	Soils	09/12/06	10/03/06		
3rd	Plant Growth & Development			02/06/07	02/27/07
3rd	Human Body	09/20/06	10/11/06	01/31/07	02/21/07
3rd	Investigating Objects in the Sky*			02/13/07	
4th	Magnetism & Electricity	09/12/06	10/03/06	01/31/07	02/21/07
4th	Food Chemistry			02/01/07	02/22/07
4th	Rocks & Minerals			01/31/07	02/21/07
4th	Animal Studies	09/07/06	09/28/06	02/08/07	03/01/07
5th	Weather Systems	09/12/06	10/03/06		
5th	Landforms	09/19/06	10/10/06	02/06/07	02/27/07
5th	Motion & Design			01/30/07	02/20/07
5th	Ecosystems	09/21/06	10/12/06		
6th	Solar System*	09/13/06			
6th	Earth's Crust*	09/19/06			
6th	Energy Transfer & Transform.*	11/16/06			
6th	Cycling of Matter & Pop. Dyn.*			02/07/07	
7th	Thrill Ride	09/20/06	10/11/06	02/08/07	03/01/07
7th	Weather & Water	09/19/06	10/10/06	02/01/07	02/22/07
8th	Microbiology/Cell Theory*	08/31/06			
8th	Hydrology*	11/02/06			
8th	Chemistry*			01/11/07	
8th	Evolution*			03/29/07	

* 1 day training

Figure 14. TASC 2006-2007 Training Schedule

On August 28, 2006, the TASC project director emailed to this researcher his revisions to these four tests:

> Attached are the test files for Human Body, Ecosystems, Investigating Weather Systems, and Landforms with our revisions in "track changes." I took care to maintain the position of the correct answer. Commentary on each is below.
>
> *Human Body Grade 3*
> 1a) Question #1 Norm & I reworked item #1. The problem appeared to be that too many answered correct on the pre because the skull is so obviously a protective bone, so we reduced all of the choices to single choices and eliminated the skull. We also tried to make the stem more concise.
> 1b) Question #6 Norm thought that the question tests what we want to know and should not be changed, worded as is. Question #6 tests content from the kit taught directly, as tested. The item did respond to instruction, but 50% of students did not answer correctly. We think that a significant portion of teachers did not teach the topic but that the item needs to remain unchanged.
> 2) The test does match the blueprint
> 3) The test matches the proportion of time spent on each subject.
> 4) We believe that 10 items is sufficient on this test.
>
> *Ecosystems*
> 1) Question #5 - the question deals with experimental design, covered in the kit and instruction to teachers, but we felt it was too wordy and difficult for 5th graders, which was why it didn't respond to instruction (slightly negative).
> 2) Test matches blueprint
> 3) Test topic distribution matches proportion of time spent on topics in instruction.
> 4) 10 items is sufficient.
>
> *Investigating Weather Systems*
> 1a) Question #1 - deleted.
> 1b) Question #2 (now #1) - choice b was a technical distinction that students could miss and mentioning the water cycle was a distraction. We want to know if students understand that cloud formation is the result of water vapor condensing to become liquid water droplets. So, we changed both stem and answer choices.
> 1c) Question #3 (now #2) students were thrown off either by seeing the earth lit or by the north pole label, so we made all of the earths identical. The item should test position only. Sunlight and shadow on the earth is an unnecessary distraction. Labeling north and south on the globe should be separate from the diagram.

(Would it be better in the stem? e.g. "In the diagram, the north pole points toward the top of the page.")
1d) Question #4 (now #3) students may again have been thrown off by seeing sunlight and shadow on the earth (now removed). We are testing whether students/teachers understand that it's winter in the southern hemisphere when the N. pole points toward the sun. Hence, N. pole now labeled.
1e) Question #5 (now #4), I suspect that because most students chose "d," they just looked at the picture and thought it looked hot, so no breeze was blowing. We have eliminated that choice.
1f) Question #6 (now #5), again we are trying to get students to correlate directness of sun's rays with temperature. The shadow on the earth was a distraction. The equator helps them see the directness of the rays, so we put it in.
1g) Question #7 (now #6) should come before question #5 (Terry, can you move it? It has already made the question numbering somewhat confusing since I deleted question #1) I simplified the wording and removed one of the choices to make them all parallel in construction.
1h) Question #8 Deleted - already covered by Question #2 (now #1)
1i) Question #9 (now#7) - unchanged
1i) Question #10 (now #8) reworked to make more parallel
1j) Question #11 (now #9) reworded slightly
1k) Question #12 Deleted
1l) Question #13 (now #10) should remain unchanged, fine as-is. Teachers probably did not teach this one. It's direct from the kit
1m) Question #14 (now #11)
1n) Question #16 (now #13) performed OK but was technically inaccurate. There is a thermosphere where it does temporarily get a little warmer, but that's above where a balloon would go. So, I reworded it and deleted a choice to keep the choices more parallel.
2) Test matches blueprint better with item #1 deleted. Deleting item #8 did not harm the match to the blueprint.
3) Very little time spent on question #1 (now deleted), but otherwise test topic distribution matches of time spent on topics in instruction.
4) 14 items are sufficient.

Landforms
1a) #2, "deltas" was confusing the kids so we took it out.
1b) #4 and #5 are reversed. We replaced the diagram and added creek names, thereby simplifying.
1c) #6 reworded and added another choice
1d) #7 improved wording to avoid confusing students
1e) #9 reversed first two choices and reworded them to clarify
1f) #10 and #11 deleted
1g) #14 (now #12) deleted

2) Test matches blueprint (deletions actually made the match slightly better)
3) Test topic distribution matches time spent on topics in instruction.
4) 12 items is sufficient.

On August 29, 2006, this researcher emailed her response to the TASC project director indicating that his changes looked fine. In addition, she asked him to indicate by underlining the correct answer for each item and to verify the order of the choices. She requested he email any additional changes and/or to let her know that the tests were ready for final formatting.

On August 30, 2006, the TASC project director emailed to this researcher the following questions and comments:

> With regard to what you said about logical order of answer choices, I wondered if that extends to a logical order for questions as well. For example, on the Ecosystems test, if question 10 came before question 6, students would have a better chance of predicting the likely outcome in question 6. I don't remember of that is desirable or undesirable so I didn't make the change. If you think it's useful, please do make that change.
>
> In the IWS test, to an alert student, question #4 fairly directly gives away the answer to question #6, with or without having studied the unit. This might have contributed to 35% of students getting it right on the pre-test.
>
> In the IWS test, I further clarified the stem of question #3 to assure that the diagram could not be misinterpreted.

In this researcher's August 31, 2006, response to the TASC project director, she indicated that one test question should not help a test-taker correctly answer another question and that one of the two questions should be changed. In addition, this researcher

included as attachments to her email “Test-taking directions to be read to teachers” and “Test-taking directors to be read to students."

On September 1, 2006, the TASC project director emailed to this researcher the following comments, after rethinking his August 30 comments:

> I've looked at Ecosystems test questions 6 and 10 again. I've changed my mind about them. They don't need to be changed. The two questions ask about different things and #10 does not help the test-taker answer question #6 (or vice versa). #6 is about watering soil with water that already contains dissolved nutrients, and it speaks to the effect of nutrients on the aquarium water. Students would answer question 6 based on experience with the "ecosystem" bottles. #10 is about a different issue, which is the fact that water, on its way through the soil, dissolves and carries nutrients into water supplies. It is not about the effects of those nutrients once they're in the aquarium. So, I was wrong and we should leave this test as-is.
>
> Regarding IWS, Q4 did in fact give away Q6. This can be easily corrected by changing the stem to Q4. The Q4 stem read: In the picture below, it is mid-morning and the sun has heated the land, making it warmer than the water. Which way is the wind probably blowing?
>
> I changed it to read: In the picture below, the land is warmer than the water. Which way is the wind probably blowing?
>
> Without time of day mentioned, the Q4 stem still provides what's needed to know that a sea breeze would be blowing but gives no information about land heating more rapidly than water. As corrected, the question doesn't help answer Q6, about soil heating more than water over a given time period. They're both related to #10, but don't give it away either, because it is asking students to draw specific conclusions from a graph.
>
> I've attached the correct IWS test with the change in "track changes."

This researcher, on September 4, 2006, responded to the TASC project director that the Investigating Weather Systems test "looks good", but that the reading level of item #6 on the Ecosystems test was at a twelfth grade reading level. This researcher sent

a revised item with a simpler sentence structure "without compromising the integrity of the question."

In three separate September 4, 2006, emails to the TASC project director, this researcher sent the following files:

- Revised Grade 5 Investigating Weather Systems test for printing,
- Revised Grade 5 Landforms test for printing, and
- Revised Grade 3 Human Body test for printing.

In addition, she sent the "assessment pack" memo that explained what each "pack" should contain, TASC to teachers directions, and teachers to students directions.

On September 5, 2006, this researcher and the TASC project director, through iterative emails, continued their work on Ecosystems item #6. This test was finalized and a separate email to the TASC project director included the revised Ecosystems test along with the memo to teachers and with directions. This completed the revisions to the tests needed for the fall 2006 test administrations.

On September 6, 2006, this researcher emailed to the TASC project director a few comments regarding test administration:

> For the pretesting and posttesting of teachers, I do not think that TASC instructors should administer the tests to the teachers. The reason for this is that not everyone agrees with the importance of these tests, and I believe this gets communicated (one way or the other) to the teachers. . . .
>
> For the pretesting and posttesting of students, I strongly believe that students' TASC-trained teachers should administer the tests. My reason is that having their own teacher administer the tests may reduce students' level of stress and/or may motivate them to do better than they may do if someone unknown to them administers the test. . . .

> As a reminder, the teacher ID number must be on their students' pretests and posttests. Otherwise, I have no way to match students to teachers/classrooms. This was spelled out in the Tchr to Stdts directions emailed to you yesterday.

In his response of the same date, the TASC project director stated:

> I agree that non-TASC instructors should administer the tests to the teachers for the reasons you gave ….
>
> I agree, for the reasons you gave plus others of my own, that TASC-trained teachers should administer the pre- and post-tests to their own students. . . .

In addition to the revisions to these four tests, TASC and CERE had been trading emails concerning a new statement of work, including budget, that would begin October 1, 2006 and end either August 31, 2007 or September 30, 2007. In a September 11, 2006, email, the TASC project director stated:

> This is to follow the brief phone contact we had this afternoon. As I said, for the past few days, I've been going over your budget, your SOW, and the TASC budget trying to figure out where to find money and what to do. <u>After going over the books, we've determined that we can't afford what you've proposed</u>. We are very pleased with your work and have enjoyed working with you, but <u>our budget constrains us to ask you to complete development of the Motion & Design Test to the level of the other 4 tests, and then request no further work</u>. If you're willing, please submit an SOW and budget for just for development of the Motion & Design Test to the level of the other 4 tests.

By October 4, 2006, a one-month extension to the existing subcontract had been granted so that this researcher could complete the work on the Motion & Design test "to the level of the other 4 tests." On October 4, 2006, this researcher emailed the TASC project director the following:

> To complete the work on the Motion & Design tests, below are TASC's action items, taken from CERE's Final Report:
>
> Gr 5 Motion & Design--Form A
> - Items 1 through 7 and 9 through 13 should be examined and reworked (or replaced).
> - Does the test match its blueprint?
> - Does the percentage of time spent on each objective need to be revised?
>
> Gr 5 Motion & Design--Form B
> - Items 1, 3, and 6 through 13 should be examined and reworked (or replaced).
> - Does the test match its blueprint?
> - Does the percentage of time spent on each objective need to be revised?
>
> Let me know what revisions you would like me to incorporate in these tests (assuming TASC wants to retain both forms).

The TASC project director's response to this researcher was that he expected to send revisions by the next day and that he assumed "we'll combine Form A and Form B to come up with a single test."

On October 9, 2006, the TASC project director emailed the following to this researcher:

> The Motion & Design tests took a fair amount of revising because, in both form A and B, the proportion of items v. proportion of time spent in instruction on objectives did need to be substantially revised. As a result, to ease your job on this, I combined all of the items I had reworked from both forms A and B into a single test, the attached file titled "RvsdGr5MDtst.A&B.cmbined.doc"
>
> I also worked to better match the test to the blueprint on form A and B. I might have succeeded but am not sure.
>
> As you'll see, on Form A, I did rework or replace items 1 through 7 and 9 through 13, I also revised #8, because although it performed well, it could have been clearer. You'll also see that on Form B, I reworked or replaced items 1, 3, and 6 through 13.

Please let me know what you think and what needs further work.

On October 11, 2006, this researcher responded to the TASC project director with the following general comments:

- Within a question, be consistent with the use of the terms "vehicle" and "car." My suggestion is that if "vehicle" and "car" are synonymous, use the simpler term—car.
- Replace numbers written as words with numbers written as figures; I think using figures is easier for students to understand—e.g., 1 or 2 instead of one or two.
- Please identify which instructional objective(s) each question is assessing and which thinking skill is being required of the test-taker.
- What do you want students to be know and be able to do—read well and understand the nuances of language, or be able to demonstrate their knowledge and understanding of laws of physics? Assuming it is to demonstrate knowledge/understanding of physics, language should not be allowed to get in the way. With a few exceptions, I think language gets in the way on most of these questions.

This researcher then provided feedback for each of the 15 items on the two test forms.

For example, she wrote:

Question 3
. . . What precisely is it that you want students to know and be able to do to answer this question correctly? Is the stem providing the information—clearly and free of extraneous information—students need to answer this question correctly?

As for the choices a, b, c: (a) is the shortest and (c) is the longest. I suspect that that is how students will select their choice especially if they are confused by the question. . . .

Question 4
I think the choices could be made more parallel in form, e.g.:
a) Increased distance traveled by using larger wheels.
b) Decrease friction by reducing the number of wheels.

c) Increase friction by causing something to rub against the wheels or the table top.
d) Decrease weight by removing some parts of the vehicle to make it less heavy.

Question 5
- Even without reading the question, I'd pick (a) because it's the longest answer.
- The choices should be made more parallel.

Question 6
Would another "balanced" choice be appropriate? I.e., "The two forces are balanced, so the wheel turns and the vehicle remains still."

Question 7
- Actually, examinees would not even need to read the stem to answer this question. All they need is the table.
- How plausible is choice (d)?

Question 10
Is (c) plausible?

Question 11
- Even without reading the question, I'd pick (b) because it's the longest answer. It is also the only choice with a reason—"that turns the back wheels"

This researcher and the TASC project director continued their iterative work by email on the Grade 5 Motion & Design test from October 19 through October 23, 2006. The final version of the test was emailed by this researcher to the TASC project director on October 23, 2006.

Factors that Affected Pilot Test Revision

There were no factors that adversely affected the test revision task, per se. Of the phase 3 tasks, test revision was the most problem-free. Working through iterative emails, the TASC project director and this researcher revised five of the seven tests that required revisions: Grade 3 Human Body, Grade 5 Investigating Weather Systems, Grade 5

Landforms, Grade 5 Ecosystems, and Grade 5 Motion & Design. The subcontract itself, or rather its completion and non-renewal, affected the two remaining third grade tests (i.e., Plant Growth & Development and Investigating Objects in the Sky) in that they were not revised, at least not under CERE's subcontract with TASC.

To summarize, the predominant factor that affected phase three was TASC itself—its scientists and its participating teachers. Almost all tasks within phase three were affected by either or both the scientists and the participating teachers. The teacher-item writer recruitment was affected not only by TASC's gross underestimation of the time required to recruit teacher-item writers but by its assumption that participating teachers would be understanding and responsive to TASC's need for pre- and post-test data that could be matched by student and teacher. Instead, TASC assumed that its participating teachers would *want* to attend a three-hour item writing workshop and write questions that would be administered to them and their students.

Item generation and revision were also affected by assumptions made by the TASC project director. One assumption he made was that teachers, once trained on a curriculum unit, would use that unit in their classrooms and, therefore, would know the science content well enough to write test questions. He also assumed that the teachers, after receiving one half-day of item writing training, would be able to write items of "medium to hard" difficulty and required higher level thinking skills. Finally, the TASC project director assumed that compensation and renewal credits would be sufficient incentives for teachers to complete the item-writing task. As previously documented, all these assumptions were shown to be overly optimistic. The teachers did not necessarily

use the curriculum units in their classrooms even though they had received the TASC training. The teachers did not know the science content well enough, even with a half-day item writing workshop, to write higher order thinking skills questions. With 60 percent of the initial item-writers dropping out, clearly compensation and renewal credits, while valuable incentives to the teachers, were not enough.

Additionally, teachers' lack of content knowledge and item writing skills caused a huge delay in receiving final items from the remaining teacher-item writers that affected pilot test assembly. Three weeks had been allotted for the item-writing task; instead, it took eight weeks. Even though additional time was provided, the quality of the items was considered poor by the content experts (i.e., the scientists)—a huge disappointment to the TASC project director who thought the Training Institute teachers should have been "the cream of the crop" as item-writers. This extension of time given to the item-writers resulted in an extraordinarily compressed timetable for test assembly, in a few cases, less than a week to have a test ready to be piloted. This compressed timetable, in turn, resulted in the elimination of the two versions—teacher and student—for each test. In addition, this compressed timetable affected the TASC scientists who then had to write items even as they prepared for the upcoming workshops they would lead. Last of all, the compressed timetable affected the printing, packaging, and delivery of the packaged tests to TASC's warehouse in time to be shipped with the science kits. However, even with the compressed timetable, out of eleven tests that were scheduled to be developed and piloted in fall 2005, three were dropped and only one pilot test deadline was missed (Grade 8 MicroLife), to be included in a February 2006 pilot test administration.

Pilot test administration was affected by reluctant participants—both scientists and teachers. While the reluctance may have stemmed from participants' lack of understanding concerning how much personal involvement they were to have in the testing process, it resulted in very low quality of the returning data, requiring extensive cleaning by CERE, from the fall 2005 test administrations. Prior to the February 2006 pilot test administrations, TASC and CERE worked together to address teachers' reluctance, and some improvement—in the quality of the test data and in the greater numbers of students for whom matching pre-post data were received—resulted.

Finally, the test revision task was the most problem-free of the entire phase three. Although seven tests required revisions, five were revised in fall 2006 before the TASC-CERE subcontract ended, after being extended one month, in October 2006 at TASC's request.

Operational Test (Phase 4)

Because TASC did not extend another subcontract to CERE, no further work—i.e., revisions to the two remaining tests and data analyses from the administrations of the operational tests—was performed by CERE for TASC under its NSF subaward.

The next chapter presents the conclusions and recommendations from this project.

CHAPTER V

DISCUSSION

This dissertation focused on the *process* of test development within an evaluation and examined planned (i.e., expected) and actual (i.e., observed) test development, specifically concentrating on the factors that affected the test development process. Planned test development was defined as the process of creating tests according to the well-established test development procedures recommended by the AERA/APA/NCME 1999 *Standards for Educational and Psychological Testing.* Actual test development was defined as the process of creating tests as it actually took place.

The project of interest—Teachers and Scientists Collaborating (TASC) at Duke University—took place within the larger context of NSF's Math-Science Partnership Program that, in turn, took place within the larger context of science education reform. As stated previously in Chapter One, TASC is merely representative of what is, most likely, common practice in the evaluation of science education programs.

Under its contract with NSF, Duke University was required to provide an evaluation of the TASC project to guide the annual assessment of progress and to measure the impact of the project's efforts. One of TASC's anticipated outcomes was improvement in teachers' and students' science content and process knowledge. To evaluate the achievement of this outcome necessitated the development of science tests to

measure improvements in content knowledge of participating teachers and their students. As test developer for TASC's project evaluation, the goal was the creation of these tests—each test to be aligned with one of the science competency goals of the 2004 NC Standard Course of Study for third, fifth, and eighth grade—according to established measurement procedures in the field of psychometrics.

Because case study provides an in-depth, longitudinal examination of an event (i.e., case), it was selected as the appropriate methodology to examine this difference between planned and actual test development. The case (or unit of analysis) was the test development task, a task that was bounded by the context in which it occurred—and over which this researcher had no control—and by time. The purpose for studying the case was to gain a more in-depth, holistic understanding of the real-life test development task that took place within a project evaluation context. In particular, this case study investigated how the actual test development process was affected by:

6. the national and state (i.e., NC) science standards,
7. the NSF's definition of "evidence" in a project evaluation,
8. the MSP project's understanding of the role of the to-be-developed tests in their project evaluation,
9. the MSP project's understanding of the test development process, and
10. the MSP project's participants (e.g., teacher item-writers and scientists).

Chapter Four documented the planned test development, including the contractual basis for developing the tests and the steps and standards followed by this researcher as test developer, the actual test development process, and the factors that affected each

phase of the test development process. These results are summarized in the following section.

Summary of Results

Factors that Affected Phase 1(Test Framework) of the Test Development Process

Factors that affected the test framework—the purpose of the test and the scope of the construct—included:

- the national and state science standards: The *National Science Education Standards* affected the scope of the construct in this phase of test development through its impact on the goals of the Math Science Partnership (MSP) program and, through its standards of science content, the content of instruction. In 2004, the NC Department of Public Instruction (DPI) incorporated into its revised Standard Course of Study for Science these standards that articulated what students should know and be able to do and that emphasized the *use* of scientific knowledge, rather than mere recitation.

  TASC, one of the grants awarded under NSF's MSP program, provided professional development training to participating teachers on science units aligned with grade-specific science competency goals of the 2004 NC Standard Course of Study.

- the NSF's definition of "evidence" in a project evaluation: Even though NSF cited the use of qualitative and quantitative data sources as appropriate evidence of effectiveness in project evaluations, this researcher documented in Chapter Four that NSF, for funding purposes, preferred quantitative "evidence." Thus, when a MSP project proposed improvement in science content knowledge, the most-highly-

favored-by-NSF evidence it could provide was student and teacher test data because NSF frequently used changes in test performance to determine whether a project had been successful in accomplishing its goals and objectives. In response to NSF's preference for quantitative "evidence," TASC, in its evaluation plan, proposed the development of science content tests.

- the TASC project personnel's understanding of the role the tests were to play in the evaluation of their project: In its proposal to NSF, TASC articulated its understanding of the role of the to-be-developed tests. TASC proposed the development of *two* sets of tests:
  - short curriculum unit topic content tests that would be used to evaluate changes in teacher and student science content knowledge, and
  - large assessment state tests, designed to authentically measure student performance in science, that would be used to evaluate whether the project met its annual milestone; that is, the observation of a statistically significant positive difference between students of participating and non-participating teachers in partner districts.

Factors that Affected Phase 2 (Test Specifications) of the Test Development Process

Factors that affected the test specifications phase of the test development process included:

- the national and state science standards: These standards emphasized the active nature of science instruction—i.e., inquiry-based—that was central to TASC's teacher training. This inquiry-based emphasis affected the tests' blueprints in that the tables

of specifications were to reflect both NC SCS instructional objectives *and* NC thinking skills, particularly higher level thinking skills (i.e., no knowledge-level items), and that test questions were to be of "medium to hard" difficulty reflecting the *use* of science content knowledge.

- TASC's understanding of the role of the to-be-developed tests in their project evaluation: TASC's initial understanding—articulated in its project evaluation, in its implementation plan document, and in its subcontract with NC DPI—was that TASC expected to use test results from short curriculum unit content tests and from large assessment state science tests as quantitative evidence to its funding agency (NSF) that TASC was meeting its stated teacher and student outcome goals.

  As documented in Chapter Four, this initial understanding of the role of the to-be-developed tests appeared to change somewhat by 2005 when TASC subcontracted CERE to develop separate tests—ten for students and ten for teachers—on science content and science process tied to the NSF-approved curriculum units that were tied to the NC SCS. Even though the purpose of the tests had not changed from TASC's original intentions as stated in its evaluation plan—that is, to measure improvements in content knowledge of participating teachers and their students, the use of the test results apparently *had* changed from providing quantitative evidence that TASC was meetings its teacher and student outcome goals to providing "evidence from [workshop] training".

- TASC's understanding of the test development process: TASC documents (proposal, evaluation plan, implementation plan, 2004 subcontract with DPI) articulated what

appeared to be TASC's understanding of test development. That is, TASC appeared to understand that:

- the tests items needed to be aligned with instruction (i.e., NC Standard Course of Study and the curriculum units);
- measurement personnel (i.e., DPI) and content experts (i.e., TASC scientists) needed to work together to insure that the tests would be content valid;
- the tests needed to be reliable;
- the tests needed to be piloted before being used operationally;
- measurement personnel were needed to analyze data from the piloted tests;
- test development was iterative with results from pilot testing potentially impacting the operational tests; and
- developing tests takes time.

Based on TASC's apparent understanding of the test development process, the expectations were that the development of the content framework would have been very systematic (it was not), that sufficient time would have been allotted for the development of ten operational tests (it was not, due in large part to the two-year delay caused by the previous test developer), and that the implicit assumptions about item writers were valid (they were not).

Factors that Affected Phase 3 (Construction, Administration, and Evaluation of Pilot Tests) of the Test Development Process

The predominant factor that affected phase three was TASC itself—its scientists and its participating teachers. Almost all tasks within phase three were affected by either

or both the scientists and the participating teachers. The teacher-item writer recruitment was affected not only by TASC's gross underestimation of the time required to recruit teacher-item writers but by its assumption that participating teachers would be understanding and responsive to TASC's need for pre- and post-test data that could be matched by student and teacher. Instead, TASC assumed that its participating teachers would *want* to attend a three-hour item writing workshop and write questions that would be administered to them and their students.

Item generation and revision were also affected by assumptions made by the TASC project director. One such assumption was that teachers, once trained on a curriculum unit, would use that unit in their classrooms and, therefore, would know the science content well enough to write test questions. A second assumption was that the teachers, after receiving one half-day of item writing training, would be able to write items of "medium to hard" difficulty and required higher level thinking skills. A third assumption was that compensation and renewal credits would be sufficient incentives for teachers to complete the item-writing task. Chapter Four documented that these assumptions were overly optimistic. The teachers did not necessarily use the curriculum units in their classrooms even though they had received the TASC training. The teachers did not know the science content well enough, even with a half-day item writing workshop, to write higher order thinking skills questions. With 60 percent of the initial item-writers dropping out, clearly compensation and renewal credits, while valuable incentives to the teachers, were not enough.

Additionally, teachers' lack of content knowledge and item writing skills caused a huge delay in receiving final items from the remaining teacher-item writers that affected pilot test assembly. Three weeks had been allotted for the item-writing task; instead, it took eight weeks. Even though additional time was provided, the quality of the items was considered poor by the content experts (i.e., the scientists)—a huge disappointment to the TASC project director who thought the Training Institute teachers should have been "the cream of the crop" as item-writers.

This extension of time given to the item-writers resulted in an extraordinarily compressed timetable for test assembly, in a few cases, less than a week to have a test ready to be piloted. This compressed timetable, in turn, resulted in the elimination of the two versions—teacher and student—for each test. In addition, this compressed timetable affected the TASC scientists who then had to write items even as they were preparing for the upcoming workshops they would lead. Finally, the compressed timetable affected the printing, packaging, and delivery of the packaged tests to TASC's warehouse in time to be shipped with the science kits. However, even with the compressed timetable, out of eleven tests that were scheduled to be developed and piloted in fall 2005, three were dropped and only one pilot test deadline was missed (Grade 8 MicroLife), to be included in a February 2006 pilot test administration.

Pilot test administration was affected by reluctant participants—both scientists and teachers. While the reluctance may have stemmed from participants' lack of understanding concerning how much personal involvement they were to have in the testing process, it resulted in very low quality of the returning data, requiring extensive

cleaning on CERE's part, from the fall 2005 test administrations. Prior to the February 2006 pilot test administrations, TASC and CERE worked together to address teachers' reluctance, and some improvement—in the quality of the test data and in the greater numbers of students for whom matching pre-post data were received—resulted.

Finally, the test revision task was the most problem-free of the entire phase three. Although seven tests required revisions, five were revised in fall 2006 before the TASC-CERE subcontract ended, after being extended one month, in October 2006 at TASC's request.

Factors that Affected Phase 4 (Construction, Administration, and Evaluation of Operational Tests) of the Test Development Process

No factors affecting this phase were documented because TASC did not extend another subcontract to CERE for the revisions of the two remaining tests and data analyses from the administrations of the operational tests.

Conclusions and Recommendations

Evaluators are in the business of "collecting, analyzing, interpreting, and communicating information about the workings and effectiveness of social programs" (Rossi, et al., 2004, p. 2). When content knowledge acquisition is a primary objective of educational programs, evaluators frequently use tests to measure this objective.

However, as pointed out by Rossi, et al., (2004), when measures must be developed to appraise a project's outcomes of interest, frequently, there is rarely sufficient time and resources to do this properly within the evaluation. These authors acknowledged that constructing such measures so that they consistently measure what

they are supposed to measure is often not easy. In addition, Wolf and Cumming (2000) commented that there was "remarkably little discussion in the academic literature" as to how an instrument actually gets developed. These authors determined, through their development of an assessment instrument, that test construction was anything but routine and unproblematic, as implied by the field of psychometrics.

The results from this case study confirm that constructing psychometrically sound measures within an evaluation is not easy, that sufficient time and resources to construct such measures properly are seldom provided, and that test construction—at least within an evaluation—is not routine and unproblematic.

Even though project directors, project evaluators, and NSF program managers may know *about* test development, they may not necessarily know the steps, and/or standards, by which tests are created. This unfamiliarity may result, as it did in this case study, in a project director allocating insufficient time and resources to develop psychometrically sound assessments and in making unrealistic assumptions about teacher-participants, item-writing, and pilot test administration. Unfamiliarity with the creation of psychometrically sound assessments may also result in a test being used before it had been properly developed, as occurred in this case study when the evaluator administered one of the tests before it had been piloted and analyzed.

In addition, unfamiliarity may result in a program manager's lack of scrutiny of an evaluation plan proposing the development of content tests. This case study documented NSF's strong preference for quantitative data in project evaluations and that NSF uses changes in test performance to determine whether a project successfully accomplished its

goal of improvement in science content knowledge. A project unable to demonstrate such improvement may forfeit its NSF funding, or a proposal failing to include the use of such data may be rejected outright. With such high stakes—at least from the perspective of a project or proposed project—placed upon the use of test data, it would seem that NSF program managers should be familiar with how psychometrically sound tests are developed in order to more effectively judge evaluation plans proposing the development of content tests.

- Recommendation 1: MSP project directors should be familiar with the steps and standards used to develop psychometrically sound tests so they can effectively judge the time and resources necessary to develop psychometrically sound assessments.
- Recommendation 2: Project evaluators should be familiar with the steps and standards used to develop psychometrically sound tests so they can effectively judge the appropriateness of a test to measure the outcome of interest (e.g., science content knowledge) and/or so they do not inadvertently circumvent the test development process where tests are to be created within the evaluation.
- Recommendation 3: NSF program officers should be knowledgeable about the development of psychometrically sound tests in order to more effectively judge evaluation plans proposing the development of content tests.
- Recommendation 4: In keeping with its requirement for evidence-based design and outcomes, NSF should outline centralized test development procedures in its program solicitations (i.e., requests for proposals).

- Recommendation 5: A future research area may be a meta-analysis of all other MSPs, examining *only* the test development process.

This case study also raises the issue of conflicting priorities that may exist within a project. This case study documented that the priority of the TASC trainers was to provide teachers with enough science content to make them feel comfortable with using the science units in their classrooms thereby incorporating a more inquiry-based approach to teaching science, an approach supported by the *NSES* science teaching standards. The evaluator's priority, on the other hand, was to provide quantitative evidence to NSF of the project's effectiveness by demonstrating a 20 percent increase in science content knowledge between pre- and post-testing in a two-day workshop (separated by about three weeks). The evaluator's priority to provide quantitative evidence to NSF resulted in one of the tests being used before it had been piloted and analyzed, which conflicted with the test developer's priority to create psychometrically sound assessments.

- Recommendation 6: In addition to an understanding of the time and resources required to develop psychometrically sound assessments, MSP project directors should understand the importance of identifying all stakeholders —such as teachers and/or project scientists—and including them in the project evaluation process.

In May 2007, the U.S. Department of Education published the *Report of the Academic Competitiveness Council* (ACC). The ACC, established by the *Deficit Reduction Act of 2005* (P.L. 109-171), was charged to:

- Identify all federal programs with a mathematics or science education focus;

- Identify the effectiveness of those programs;
- Determine areas of overlap or duplication among those programs;
- Identify target populations served by such programs; and
- Recommend processes to efficiently integrate and coordinate those programs.

In addition to its legislated responsibilities, the ACC also set milestones to guide its mandate:

- Delineate the goals of the programs;
- Determine the extent to which the programs have undergone independent, external evaluation based on sound, scientific principles;
- Ascertain the extent to which the programs have quantitative evidence of achieving their stated goals;
- Establish standards for measuring and evaluating these programs, including common measures as appropriate; and
- Formulate recommendations for administrative or legislative action that, if carried out, would more efficiently integrate and coordinate federal spending on STEM education programs.

Three working groups were formed—K-12 Education, Postsecondary Education, and Informal Education and Outreach—and each group developed common goals and measures for its programs. The K-12 Education working group adopted goals and measures at both the national level and the program and project level that focus primarily on improving student achievement, teacher quality, and student engagement.

To determine STEM program effectiveness, the ACC used a Hierarchy of Study Design (Figure 15), proposed by the Coalition for Evidence-Based Policy, an independent organization whose mission is to promote government policymaking based on rigorous evidence of effectiveness.

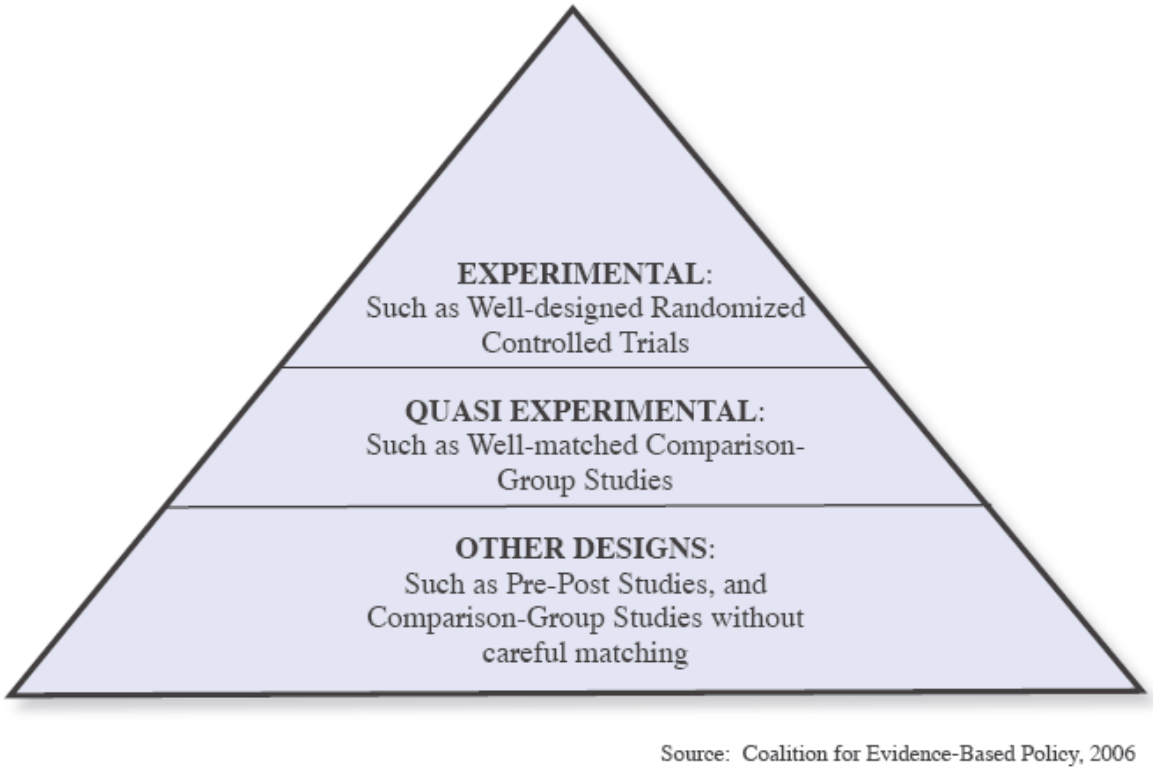


Figure 15. Hierarchy of study designs for evaluating the effectiveness of a STEM educational intervention, by expected distribution of study type

In its reported findings from the assessment of program effectiveness of all the federal agencies with education programs aimed at improving America's competitiveness in science, technology, engineering, and mathematics (STEM), the ACC identified a total of 24 STEM programs primarily focused on elementary and secondary education

outcomes, administered by eight federal agencies, with total fiscal year 2006 funding of approximately $574 million. Of these 24 programs, NSF administered the greatest number, with six programs, and provided the largest amount of funding, i.e., $242 million, $63.2 million (or 26 percent) of which was provided for its MSP program. The U.S. Department of Education came in second at $239 million in program funding, $182 million (or 76 percent) of which was allotted for MSPs. Together, NSF and DOE accounted for approximately 85 percent of the total K-12 education STEM program funding.

One of the ACC's key findings from its K-12 Education working group was that many agencies have judged their funded programs on the basis of inputs (i.e., number of teacher-participants) or on surveys of attitudinal changes, or have concluded that certain programs were effective based on their management processes. The *Report* stated that "a more appropriate method to measure educational impacts is to assess outcomes, the most direct indicators of effectiveness, and require programs to adopt consistently high standards for determining and comparing their impact." (p. 22). This statement further supports the need for psychometrically sound assessments to measure proposed educational outcomes.

In a second key finding, the K-12 Education working group indicated that many of the K-12 goals and measures adopted by the ACC (e.g., percentage of students scoring at or above proficient on state, national, and international science assessments) align with the expectations now set for the nation's K-12 education system under *No Child Left Behind* (NCLB). It further indicated that the school accountability framework in *NCLB*

presents an important opportunity for enhancing our understanding of the impact of federal science and math education programs. The *Report* stated:

> Potentially, the federal government can learn whether many of the science and math education activities it supports yield measurable student achievement gains through existing assessment activities in local school systems. . . .
>
> Where federal efforts to improve STEM education are aligned with state standards, state assessments can be used to measure the impact of federally supported activities on student learning. In these cases, scientifically rigorous impact evaluations involving randomized controlled trials or well-matched comparison groups can be carried out at reasonable cost, providing valuable information to determine whether federally supported projects are having a causal effect on student achievement. In programs where federal efforts have not been aligned with state standards, a choice must be made whether to align them or whether to adopt customized assessments to measure student learning, which may be more costly.
> (U.S. Department of Education, 2007, p. 23; emphasis added.)

In its third (and final) key finding, the K-12 Education working group indicated that "both project directors and federal program managers can use the results of student assessments to refine activities and enhance their impact on student learning" (p. 23, emphasis added).

The ACC clearly recognizes that creating customized assessments to measure student learning may be more costly than using existing state assessments. However, the implicit assumption is that state assessments are aligned with the content of instruction, which Porter (2002) asserts plays a crucial role in determining gains in student achievement.

- Recommendation 7: A future research area may be to establish a protocol that provides a systematic means by which to examine an existing or proposed MSP

project for alignment with state science standards. Such a protocol would be cost-effective in that demonstrated alignment with state science standards would enable projects to use existing state science assessments, which must be in place, according to *NCLB*, by the 2007-2008 school year, to demonstrate student achievement. In this way, project directors and evaluators, typically with limited familiarity with the steps and standards by which psychometrically sound assessments are created, would not be placed in the role of test developer.

REFERENCES

Allen, M. J. and Yen, W. M. (1979). *Introduction to measurement theory*. Prospect Heights, IL: Waveland Press, Inc.

American Association for the Advancement of Science (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology*. Washington, D.C.: author.

American Association for the Advancement of Science (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

American Association for the Advancement of Science (1995). *Project 2061: Science literacy for a changing future: A decade of reform*. Washington, D.C.: author.

American Association for the Advancement of Science (1998). *Project 2061: Science literacy for a changing future: Update 1998-99*. Washington, D.C.: author.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Educational Research Association.

American Federation of Teachers, National Council on Measurement in Education, and National Education Association. (1990). *The Standards for Competence in the Educational Assessment of Students.* (Retrieved 8/01/06 from http://www.unl.edu/buros/bimm/html/subarts.html).

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain.* New York: David McKay.

Bogdan, R. C. & Biklen, S. K. (1992). *Qualitative research for education: An introduction to theory and methods.* Boston: Allyn and Bacon.

Brookhart, S. M. (2001). *The Standards and classroom assessment research.* Paper presented at the annual meeting of the American Association of Colleges for Teacher Education, Dallas, TX. (ERIC Document Reproduction Service).

Bybee, R. W. (1997). *Achieving scientific literacy: From purposes to practices.* Portsmouth, NH: Heinemann.

Campbell, C., Murphy, J. A., and Holt, J. K. (2002, October). *Psychometric analysis of an assessment literacy instrument: Applicability to preservice teachers.* Paper presented at the annual meeting of the Mid-Western Educational Research Association, Columbus, OH.

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches.* Thousand Oaks, CA: Sage Publications.

Crocker, L. and Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Harcourt Brace Jovanovich College Publishers.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching, 37*(6), 582-601.

evaluate. (n.d.). *Dictionary.com Unabridged (v 1.0.1)*. Retrieved October 02, 2006, from Dictionary.com website: http://dictionary.reference.com/browse/evaluate

Fielding, N. G. and Lee, R. M. (1998). *Computer analysis and qualitative research.* Thousand Oaks, CA: Sage.

Flyvbjerg, B. (2006). Five misunderstandings about case study research. *Qualitative Inquiry,* 12(2), 219-245.

Fraenkel, J. R. and Wallen, N. E. (2003). *How to design and evaluate research in education, 5th edition.* New York: McGrawHill.

Fulp, S. L. (2002a). *2000 National Survey of Science and Mathematics Education: Status of elementary school teaching.* Chapel Hill, NC: Horizon Research, Inc.

Fulp, S. L. (2002b). *2000 National Survey of Science and Mathematics Education: Status of middle school teaching.* Chapel Hill, NC: Horizon Research, Inc.

Gahan, C. and Hannibal, M. (1999). *Doing qualitative research using QSR NUD*IST.* Thousand Oaks, CA: Sage.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Gullikson, A. R. (1984). Teacher perspectives of their instructional use of tests. *Journal of Educational Research, 77*(4), 244-248.

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation.* Boston: Allyn and Bacon.

Horizon Research, Inc.. (2003). The influence of the *National Science Education Standards* on teachers and teaching practice. In K. S. Hollweg and D. Hill (Eds.), *What is the influence of the National Science Education Standards? Reviewing the evidence, a workshop summary* (pp. 91-107). Washington, D.C.: The National Academies Press.

Joint Committee on Standards for Educational Evaluation (1994). *The program evaluation standards, 2nd edition*. Newbury Park, CA: Sage.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook II: Cognitive domain.* New York: David McKay.

Lee, O. (1998). *Current conceptions of science achievement and implications for assessment and equity in large education systems* (Research Monograph No. 12). Madison, WI: National Institute for Science Education, University of Wisconsin-Madison.

Marzano, R. J., Brandt, R. S., Hughes, C. S., Jones, B. F., Presseisen, B. Z., Rankin, S.C. & Suhor, C. (1988). *Dimensions of Thinking: A framework for curriculum and instruction.* Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (1992). *A different kind of classroom: Teaching with Dimensions of Learning.* Alexandria, VA: Association for Supervision and Curriculum Development.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Merriam, S. B (1988). C*ase study research in education: A qualitative approach*. San Francisco, CA: Jossey-Bass Publishers.

Merriam, S. B (1998). *Qualitative research and case study applications in education*. San Francisco, CA: Jossey-Bass Publishers.

Mertler, Craig A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education, 33*(1), 49-64.

Miles, M. B. and Huberman, A. M. (1994). *Qualitative data analysis: An expanded sourcebook*. Thousand Oaks, CA: Sage.

National Commission on Excellence in Education (1983). *A nation at risk: The imperative for educational reform.* Washington, D.C.: U.S. Government Printing Office.

National Research Council. (1996). *National Science Education Standards.* Washington, D.C.: National Academy Press.

National Science Foundation (2002a). Math and Science Partnership Program Solicitation NSF-02-061. Washington, D.C.: author.

National Science Foundation (2002b). *The 2002 user friendly handbook for project evaluation.* Washington, D.C.: author.

National Science Foundation (2005). *Math and Science Partnership Program: Strengthening America by advancing academic achievement in mathematics and science.NSF 05-069*. Washington, D.C.: author.

No Child Left Behind Act of 2001, 20 U.S.C. §6301 et. seq.

North Carolina Department of Public Instruction (2004). *Science: Standard course of study and K-12 grade level competencies*. Raleigh, NC: author.

O'Sullivan, C. Y., Lauko, M. A., Grigg, W. S., Qian, J., and Zhang, J. (2003). *The Nation's Report Card: Science 2000,* NCES 2003-453. Washington, D.C.: U.S.

Department of Education, Institute of Education Sciences, National Center for Education Statistics.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher 6*(1), 21-27.

Popham, W. J. (1999). *Classroom assessment: What teachers need to know, 2nd edition.* Boston: Allyn and Bacon.

Popham, W. J. and Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*(1), 1-9.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Punch, K. F. (2005). *Introduction to social research: Quantitative and qualitative approaches, 2nd edition.* Thousand Oaks, CA: Sage Publications.

Rossi, P. H., Lipsey, M. W., and Freeman, H. E. (2004). *Evaluation: A systematic approach, 7th edition*. Thousand Oaks, CA: Sage Publications.

Simpson, E. J. (1972). *The classification of educational objectives in the psychomotor domain.* Washington, D.C.: Gryphon House.

Smith, L. M. (1978). An evolving logic of participant observation, educational ethnography and other case studies. In L. Shulman (ed.) *Review of Research in Education.* Itasca, IL: Peacock.

Smith, P. S., Banilower, E. R., McMahon, K. C., and Weiss, I. R. (2002). *The National Survey of Science and Mathematics Education: Trends from 1977 to 2000.* Chapel Hill, NC: Horizon Research, Inc.

Stake, R. E. (1994). Case studies. In N. K. Denzin and Y. S. Lincoln (eds.), *Handbook of Qualitative Research.* Thousand Oaks, CA: Sage.

Stake, R. E. (1995). *The art of case study research.* Thousand Oaks, CA: Sage.

Stake, R. E. (1998). Case studies. In N. K. Denzin and Y. S. Lincoln (eds.), *Strategies of Qualitative Inquiry* (pp. 86-107). Thousand Oaks, CA: Sage.

State Collaborative on Assessment and Student Standards (2003). *Glossary of assessment terms and acronyms.* Washington, D.C.: Council of Chief State School Officers.

Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan, 77*(3), 238-245.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice, 18*(1), 23-27.

TASC (2002a). Proposal in response to NSF Program Solicitation 02-061.

TASC (2002b). Evaluation plan document (part of proposal to NSF Program Solicitation 02-061).

TASC (2002c). Strategic Plan EHR-0227035.

TASC (2005). Scope of Work for the Center for Educational Research and Evaluation (CERE) at the University of North Carolina-Greensboro, a Subaward under NSF Grant # EHR-0227035 to Teachers and Scientists Collaborating at Duke University.

Thorndike, R. L. (1982). *Applied psychometrics.* Boston: Houghton Mifflin Company.

Trochim, W. M.K. (1989). Outcome pattern matching and program theory. *Evaluation and Program Planning, 12*, 355-366.

Tyack, D. B. (1974). *The one best system: A history of American urban education.* Cambridge, MA: Harvard University Press.

U.S. Department of Education (2007). *Report of the Academic Competitiveness Council.* Washington, D.C.: author.

Webb, N. L. (2002). *Assessment literacy in a standards-based education setting* (WCER Working Paper No. 2002-4). Madison, WI: Wisconsin Center for Education Research.

Weiss, I. R., Banilower, E. R., McMahon, K. C., and Smith, P. S. (2001). *Report of the 2000 National Survey of Science and Mathematics Education.* Chapel Hill, NC: Horizon Research, Inc.

Wise, S. L., Lukin, L. E., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of Teacher Education, 42*(1), 37-42.

Wolf, A. and Cumming, J. J. (2000). The inside story: The reality of developing an assessment instrument. *Studies in Educational Evaluation, 26*(3), 211-229.

Worthen, B. R., Sanders, J. R., and Fitzpatrick, J. L. (1997). *Program evaluation: Alternative approaches and practical guidelines, 2nd edition.* White Plains, NY: Longman Publishers.

Yin, R. K. (2003). *Case Study Research, 3rd edition.* Thousand Oaks, CA: Sage Publications, Inc.

Zhang, Z. and Burry-Stock, J. A. (1994). *Assessment practices inventory.* Tuscaloosa, AL: The University of Alabama.

Zhang, Z. and Burry-Stock, J. A. (1995). *A Multivariate Analysis of Teachers' Perceived Assessment Competenccy as a Function of Measurement Training and Years of Training.* Paper presented at the annual meeting of the Mid-South Educational Research Association, Biloxi, MS. (ERIC Document Reproduction Service No. 393807).

Zhang, Z. and Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education, 16*(4), 323-342.

APPENDIX A

TASC TRAINING SCHEDULE, 2005-2006

		Fall Cycle Training dates		Winter Cycle Training Dates	
		Session 1	Session 2	Session 1	Session 2
K	Wood & Paper	9/7/05	9/28/05	not offered	
K	Comparing & Measuring	9/12/05	--	not offered	
K	Ant Homes Under the Ground	9/8/05	--	2/2/06	--
1st	Solids & Liquids	not offered		2/16/06	3/9/06
1st	Balance & Motion	9/12/05	--	1/31/06	--
1st	Organisms	9/15/05	10/6/05	2/9/06	3/2/06
1st	Pebbles, Sand & Silt	9/13/05	10/4/05	2/7/06	2/28/06
2nd	Changes	9/22/05	10/13/05	not offered	
2nd	Lifecycle of Butterflies	9/15/05	--	2/9/06	--
2nd	Sound	9/14/05	10/5/05	1/31/06	2/21/06
2nd	Air & Weather	9/8/05	9/29/05	2/1/06	2/22/06
3rd	Soils	9/7/05	9/28/05	not offered	
3rd	Plant Growth & Development	9/13/05	10/4/05	2/7/06	2/28/06
3rd	Human Body	9/22/05	10/13/05	2/8/06	3/1/06
3rd	Investigating Objects in the Sky	9/7/05	--	1/31/06	--
4th	Magnetism & Electricity	9/20/05	10/11/05	2/14/06	3/8/06
4th	Food Chemistry	not offered		2/8/06	3/1/06
4th	Rocks & Minerals	not offered		2/1/06	2/22/06
4th	Animal Studies	9/8/05	9/29/05	2/2/06	2/23/06
5th	Investigating Weather Systems	9/14/05	10/5/05	2/2/06	2/23/06
5th	Landforms	9/13/05	10/4/05	2/7/06	2/28/06
5th	Motion & Design	not offered		2/14/06	3/7/06
5th	Ecosystems	9/15/05	10/6/05	2/9/06	3/2/06
6th	Solar System	9/20/05	--	not offered	
6th	Energy Transfer & Transformation	not offered		2/8/06	--
6th	Cycling of Matter & Pop. Dynamics	11/1/05	--	not offered	
6th	Earth's Crust	not offered		4/4/06	--
7th	Thrill Ride	not offered		2/15/06	3/8/06
7th	Weather and Water	9/21/05	10/12/05	2/16/06	3/9/06
8th	Earth History	9/21/05	10/12/05	not offered	
8th	Micro-Life	9/20/05	10/11/05	2/15/06	3/8/06

APPENDIX B

Selected Testing Standards

Topic	Standard	Statement
Test Development	3.2	The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purpose(s) of the test and about the relation of items to the dimensions of the domain they are intended to represent.
	3.3	The test specifications should be documented, along with their rationale and the process by which they were developed. The test specifications should define the content of the test, the proposed number of items, the item formats, the desired psychometric properties of the items, and the item and section arrangement. They should also specify the amount of time for testing, directions to the test takers, procedures to be used for test administration and scoring, and other relevant information.
	3.4	The procedures used to interpret test scores, and, when appropriate, the normative or standardization samples or the criterion used should be documented.
	3.6	The types of items, the response formats, scoring procedures, and test administration procedures should be selected based on the purposes of the test, the domain to be measured, and the intended test takers. To the extent possible, test content should be chosen to ensure that intended inferences from test scores are equally valid for members of different groups of test takers. The test review process should include empirical analyses and, when appropriate, the use of expert judges to review items and response formats. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented.
	3.7	The procedures used to develop, review, and try out items, and to select items from the item pool should be documented. If the items were classified into different categories or subtests according to the test specifications, the procedures used for the classification and the appropriateness and accuracy of the classification should be documented.

Topic	Standard	Statement
	3.8	When item tryouts or field tests are conducted, the procedures used to select the sample(s) of test takers for item tryouts and the resulting characteristics of the sample(s) should be documented. When appropriate, the sample(s) should be as representative as possible of the population(s) for which the test is intended.
	3.9	When a test developer evaluates the psychometric properties of items, the classical or item response theory (IRT) model used for evaluating the psychometric properties of items should be documented. The sample used for estimating item properties should be described and should be of adequate size and diversity for the procedure. The process by which items are selected and the data used for item selection, such as item difficulty, item discrimination, and/or item information, should also be documented. When IRT is used to estimate item parameters in test development, the item response model, estimation procedures, and evidence of model fit should be documented.
	3.11	Test developers should document the extent to which the content domain of a test represents the defined domain and test specifications.
	3.19	The directions for test administration should be presented with sufficient clarity and emphasis so that it is possible for others to replicate adequately the administration conditions under which the data on reliability and validity, and, where appropriate, norms were obtained.
	3.20	The instructions presented to test takers should contain sufficient detail so that test takers can respond to a task in the manner that the test developer intended. When appropriate, sample material, practice or sample questions, criteria for scoring, and a representative item identified with each major area in the test's classification or domain should be provided to the test takers prior to the administration of the test or included in the testing material as part of the standard administration instructions.
	3.22	Procedures for scoring and, if relevant, scoring criteria should be presented by the test developer in sufficient detail and clarity to maximize the accuracy of scoring.

Topic	Standard	Statement
		Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear. This is especially critical if tests can be scored locally.
Validity	1.1	A rational should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation.
	1.2	The test developer should set forth clearly how test scores are intended to be interpreted and used. The population(s) for which a test is appropriate should be clearly delimited, and the construct [that is, the concept or characteristic that a test is designed to measure] that the test is intended to assess should be clearly described.
	1.5	The composition of any sample of examinees from which validity evidence is obtained should be described in as much detail as is practical, including major relevant sociodemographic and developmental characteristics.
	1.6	When the validation rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified in reference to the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.
	1.7	When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications, and experience, of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

Topic	Standard	Statement
Reliability and errors of measurement	2.1	For each total score, subscore, or combination of scores that is to be interpreted, estimates of relevant reliabilities and standard errors of measurement or test information functions should be reported.
	2.2	The standard error of measurement, both overall and conditional (if relevant), should be reported both in raw score or original scale units and in units of each derived score recommended for use in test interpretation.
	2.3	When test interpretation emphasizes differences between two observed scores of an individual or two averages of a group, reliability data, including standard errors, should be provided for such differences.
	2.4	Each method of quantifying the precision or consistency of scores should be described clearly and expressed in terms of statistics appropriate to the method. The sampling procedures used to select examinees for reliability analyses and descriptive statistics on these samples should be reported.
Scales, Norms, and Score Comparability	4.1	Test documents should provide test users with clear explanations of the meaning and intended interpretation of derived score scales, as well as their limitations.
	4.2	The construction of scales used for reporting scores should be described clearly in test documents.
	4.4	When raw scores are intended to be directly interpretable, their meanings, intended interpretations, and limitations should be described and justified in the same manner as is done for derived score scales.
	4.9	When raw score or derived score scales are designed for criterion-referenced interpretation, including the classification of examinees into separate categories, the rationale for recommended score interpretations should be clearly explained.
Test Administration	5.1	Test administrators should follow carefully the standardized procedures for administration and scoring specified by the test developer, unless the situation or a

Topic	Standard	Statement
		test taker's disability dictates that an exception should be made.
	5.5	Instructions to test takers should clearly indicate how to make responses. Instructions should also be given in the use of any equipment likely to be unfamiliar to test takers. Opportunity to practice responding should be given when equipment is involved, unless use of the equipment is being assessed.
	5.10	When test score information is released to students, parents, legal representatives, teachers, clients, or the media, those responsible for testing programs should provide appropriate interpretations. The interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used.
	5.13	Transmission of individually identified test scores to authorized individuals or institutions should be done in a manner that protects the confidential nature of the scores.
	5.15	When test data about a person are retained, both the test protocol and any written report should also be preserved in some form. Test users should adhere to the policies and record-keeping practice of their professional organization.
	5.16	Organizations that maintain test scores on individuals in data files or in an individual's records should develop a clear set of policy guidelines on the duration of retention of an individual's records, and on the availability, and use over time, of such data.
Supporting Documentation	6.2	Test documents should be complete, accurate, and clearly written so that the intended reader can understand the content.
	6.4	The population for whom the test is intended and the test specifications should be documented. If applicable, the item pool and scale development procedures should be described in the relevant test manuals. If normative data are provided, the norming population should be described in terms of relevant demographic variables, and the

Topic	Standard	Statement
		year(s) in which the data were collected should be reported.
	6.5	When statistical descriptions and analyses that provide evidence of the reliability of scores and the validity of their recommended interpretations are available, the information should be included in the test's documentation. When relevant for test interpretation, test documents ordinarily should include item level information, cut scores and configural rules, information about raw scores and derived scores, normative data, the standard errors of measurement, and a description of the procedures used to equate multiple forms.
	6.6	When a test relates to a course of training or study, a curriculum, a textbook, or packaged instruction, the documentation should include an identification and description of the course or instructional materials and should indicate the year in which these materials were prepared.
	6.7	Test documents should specify qualifications that are required to administer a test and to interpret the test scores accurately.
	6.13	When substantial changes are made to a test, the test's documentation should be amended, supplemented, or revised to keep information for users current and to provide useful information or cautions.
	6.14	Every test form and supporting document should carry a copyright date or publication date.

APPENDIX C

Test Blueprints

Grade 3: Plant Growth & Development			NC Thinking Skills					
Competency Goal 1: The learner will conduct investigations and build an understanding of plant growth and adaptations.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
1.01 Observe and measure how the quantities and qualities of nutrients, light, and water in the environment affect plant growth.	0%	0						
1.02 Observe and describe how environmental conditions determine how well plants survive and grow in a particular environment.	0%	0						
1.03 Investigate and describe how plants pass through distinct stages in their life cycle including. • Growth. • Survival. • Reproduction	50%	4						
1.04 Explain why the number of seeds a plant produces depends on variables such as light, water, nutrients, and pollination.	10%	1						
1.05 Observe and discuss how bees pollinate flowers.	30%	2						
1.06 Observe, describe and record properties of germinating seeds.	10%	1						

Grade 3: Soils			NC Thinking Skills					
Competency Goal 2: The learner will conduct investigations to build an understanding of soil properties.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
2.01 Observe and describe the properties of soil: • Color. • Texture. • Capacity to hold water.	15%	1						
2.02 Investigate and observe that different soils absorb water at different rates.	20%	2						
2.03 Determine the ability of soil to support the growth of many plants, including those important to our food supply.	20%	2						
2.04 Identify the basic components of soil: • Sand. • Clay. • Humus.	40%	3						
2.05 Determine how composting can be used to recycle discarded plant and animal material.	5%	0						
2.06 Determine the relationship between heat and decaying plant matter in a compost pile.	0%	0						

Grade 3: Investigating Objects in the Sky			NC Thinking Skills					
Competency Goal 3: The learner will make observations and use appropriate technology to build an understanding of the earth/moon/sun system.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
3.01 Observe that light travels in a straight line until it strikes an object and is reflected and/or absorbed.	0%	0						
3.02 Observe that objects in the sky have patterns of movement including: • Sun. • Moon. • Stars.	30%	2						
3.03 Using shadows, follow and record the apparent movement of the sun in the sky during the day.	20%	2						
3.04 Use appropriate tools to make observations of the moon.	10%	1						
3.05 Observe and record the change in the apparent shape of the moon from day to day over several months and describe the pattern of changes.	20%	2						
3.06 Observe that patterns of stars in the sky stay the same, although they appear to move across the sky nightly.	20%	1						

Grade 3: Human Body			NC Thinking Skills					
Competency Goal 4: The learner will conduct investigations and use appropriate technology to build an understanding of the form and function of the skeletal and muscle systems of the human body.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
4.01 Identify the skeleton as a system of the human body.	5%	0						
4.02 Describe several functions of bones: § Support § Protection § Locomotion	30%	3						
4.03 Describe the functions of different types of joints: § Hinge § Ball and socket § Gliding.	30%	2						
4.04 Describe how different kinds of joints allow movement and compare this to the movement of mechanical devices.	10%	1						
4.05 Observe and describe how muscles cause the body to move.	25%	2						

Grade 4: Magnetism and Electricity			NC Thinking Skills					
Competency Goal 3: The learner will make observations and conduct investigations to build an understanding of magnetism and electricity.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
3.01 Observe and investigate the pull of magnets on all materials made of iron and the pushes or pulls on other magnets.	25%	2						
3.02 Describe and demonstrate how magnetism can be used to generate electricity.	10%	1						
3.03 Design and test an electric circuit as a closed pathway including an energy source, energy conductor, and an energy receiver.	15%	1						
3.04 Explain how magnetism is related to electricity.	5%	0						
3.05 Describe and explain the parts of a light bulb.	10%	1						
3.06 Describe and identify materials that are conductors and nonconductors of electricity.	15%	1						
3.07 Observe and investigate that parallel and series circuits have different characteristics.	10%	1						
3.08 Observe and investigate the ability of electric circuits to produce light, heat, sound, and magnetic effects.	10%	1						
3.09 Recognize lightning as an electricl discharge and show proper safety behavior when lightning occurs.	0%	0						

Grade 4: Food Chemistry			NC Thinking Skills					
Competency Goal 4: The learner will conduct investigations and use appropriate technology to build an understanding of how food provides energy and materials for growth and repair of the body.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
4.01 Explain why organisms require energy to live and grow.	15%	1						
4.02 Show how calories can be used to compare the chemical energy of different foods.	10%	1						
4.03 Discuss how foods provide both energy and nutrients for living organisms.	30%	2						
4.04 Identify starches and sugars as carbohydrates.	30%	2						
4.05 Determine that foods are made up of a variety of components.	15%	2						

Grade 5: Ecosystems			NC Thinking Skills					
Competency Goal 1: The learner will conduct investigations to build an understanding of the interdependence of plants and animals.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
1.01 Describe and compare several common ecosystems (communities of organisms and their interaction with the environment).	20%	1						
1.02 Identify and analyze the functions of organisms within the population of the ecosystem: • Producers. • Consumers. • Decomposers.	20%	1						
1.03 Explain why an ecosystem can support a variety of organisms.	10%	1						
1.04 Discuss and determine the role of light, temperature, and soil composition in an ecosystem's capacity to support life.	20%	2						
1.05 Determine the interaction of organisms within an ecosystem.	10%	1						
1.06 Explain and evaluate some ways that humans affect ecosystems. • Habitat reduction due to development. • Pollutants. • Increased nutrients.	20%	2						
1.07 Determine how materials are recycled in nature	0%	0						

Grade 5: Landforms			NC Thinking Skills					
Competency Goal 2: The learner will make observations and conduct investigations to build an understanding of landforms.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
2.01 Identify and analyze forces that cause change in landforms over time including. • Water and Ice. • Wind. • Gravity.	5%	0						
2.02 Investigate and discuss the role of the water cycle and how movement of water over and through the landscape helps shape land forms.	10%	1						
2.03 Discuss and consider the wearing away and movement of rock and soil in erosion and its importance in forming: • Canyons. • Valleys. • Meanders. • Tributaries.	10%	1						
2.04 Describe the deposition of eroded material and its importance in establishing landforms including: • Deltas. • Flood Plains.	10%	1						
2.05 Discuss how the flow of water and the slope of the land affect erosion.	20%	2						
2.06 Identify and use models, maps, and aerial photographs as ways of representing landforms.	40%	3						

Grade 5: Landforms			NC Thinking Skills					
Competency Goal 2: The learner will make observations and conduct investigations to build an understanding of landforms.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
2.07 Discuss and analyze how humans influence erosion and deposition in local communities, including school grounds, as a result of: • Clearing land. • Planting vegetation. • Building dams.	5%	0						

Grade 5: Investigating Weather Systems			NC Thinking Skills					
Competency Goal 3: The learner will conduct investigations and use appropriate technology to build an understanding of weather and climate	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
3.01 Investigate the water cycle including the processes of: • Evaporation. • Condensation. • Precipitation. • Run-off.	15%	1						
3.02 Discuss and determine how the following are affected by predictable patterns of weather: • Temperature. • Wind direction and speed. • Precipitation. • Cloud cover. • Air pressure.	20%	2						
3.03 Describe and analyze the formation of various types of clouds and discuss their relation to weather systems.	5%	0						
3.04 Explain how global atmospheric movement patterns affect local weather.	10%	1						
3.05 Compile and use weather data to establish a climate record and reveal any trends.	15%	1						
3.06 Discuss and determine the influence of geography on weather and climate: • Mountains • Sea breezes • Water bodies.	35%	3						

Grade 5: Motion and Design			NC Thinking Skills					
Competency Goal 4: The learner will conduct investigations and use appropriate technologies to build an understanding of forces and motion in technological designs.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
4.01 Determine the motion of an object by following and measuring its position over time.	10%	1						
4.02 Evaluate how pushing or pulling forces can change the position and motion of an object.	20%	2						
4.03 Explain how energy is needed to make machines move. • Moving air. • Gravity.	10%	1						
4.04 Determine that an unbalanced force is needed to move an object or change its direction.	15%	1						
4.05 Determine factors that affect motion including: • Force • Friction. • Inertia. • Momentum	20%	2						
4.06 Build and use a model to solve a mechanical design problem. • Devise a test for the model. • Evaluate the results of test.	20%	1						
4.07 Determine how people use simple machines to solve problems.	5%	0						

EARTH HISTORY								
Grade 8:			**NC Thinking Skills**					
Competency Goal 5: The learner will conduct investigations and utilize appropriate technologies and information systems to build an understanding of evidence of evolution in organisms and landforms.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives								
5.01 Interpret ways in which rocks, fossils, and ice cores record Earth's geologic history and the evolution of life including: • Geologic Time Scale. • Index Fossils. • Law of Superposition. • Unconformity. • Evidence for climate change. • Extinction	50%	3						
5.02 Correlate evolutionary theories and processes: • Biological. • Geological. • Technological.	20%	2						
5.03 Examine evidence that the geologic evolution has had significant global impact including: • Distribution of living things. • Major geological events. • Mechanical and chemical weathering.	20%	2						
5.04 Analyze satellite imagery as a method to monitor Earth from space: • Spectral analysis. • Reflectance curves.	0%							

EARTH HISTORY								
Grade 8:			**NC Thinking Skills**					
Competency Goal 5: The learner will conduct investigations and utilize appropriate technologies and information systems to build an understanding of evidence of evolution in organisms and landforms.	**% of time spent on objective**	**Minimum number of items to be created**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
5.05 Use maps, ground truthing and remote sensing to make predictions regarding: • Changes over time. • Land use. • Urban sprawl. • Resource management.	10%	1						

MICRO-LIFE									
Grade 8:	**9/15/2005**	**Revised 10/6**		**NC Thinking Skills**					
Competency Goal 6: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of cell theory.	**% of time spent on objective**	**% of time spent on objective**	**No. of Items**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
Objectives									
6.01 Describe cell theory: • All living things are composed of cells. • Cells provide structure and carry on major functions to sustain life. • Some organisms are single cell; other organisms, including humans, are multi-cellular. • Cell function is similar in all living things.	30%	10%	1						
6.02 Analyze structures, functions, and processes within animal cells for: • Capture and release of energy. • Feedback information. • Dispose of wastes. • Reproduction. • Movement. • Specialized needs.	5%	20%	2						
6.03 Compare life functions of protists: • Euglena. • Amoeba. • Paramecium. • Volvox.	5%	0%	0						

MICRO-LIFE									
Grade 8:	**9/15/2005**	**Revised 10/6**		**NC Thinking Skills**					
Competency Goal 6: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of cell theory.	**% of time spent on objective**	**% of time spent on objective**	**No. of Items**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
6.04 Conclude that animal cells carry on complex chemical processes to balance the needs of the organism. • Cells grow and divide to produce more cells. • Cells take in nutrients to make the energy for the work cells do. • Cells take in materials that a	20%	15%	1						
Competency Goal 7: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of microbiology.									
Objectives									
7.01 Compare and contrast microbes: • Size, shape, structure. • Whether they are living cells.	5%	0%	0						
7.02 Describe diseases caused by microscopic biological hazards including: • Viruses. • Bacteria. • Parasites. • Contagions. • Mutagens.	10%	20%	2						

MICRO-LIFE									
Grade 8:	**9/15/2005**	**Revised 10/6**		**NC Thinking Skills**					
Competency Goal 6: The learner will conduct investigations, use models, simulations, and appropriate technologies and information systems to build an understanding of cell theory.	**% of time spent on objective**	**% of time spent on objective**	**No. of Items**	**Org**	**Appl**	**Anly**	**Gen**	**Int**	**Eval**
7.03 Analyze data to determine trends or patterns to determine how an infectious disease may spread including: • Carriers. • Vectors. • Conditions conducive to disease. • Calculate reproductive potential of bacteria.	20%	30%	2						
7.04 Evaluate the human attempt to reduce the risk of and treatments for microbial infections including: • Solutions with anti-microbial properties. • Antibiotic treatment. • Research.	5%	5%	0						
7.05 Investigate aspects of biotechnology including: • Specific genetic information available. • Careers. • Economic benefits to North Carolina. • Ethical issues. • Impact for agriculture.	0%	0%	0						

APPENDIX D

Science Item Specification Sheet

☐ **Grade 3** ☐ **Grade 4** ☐ **Grade 5**

Competency Goal:	**Instructional Objective:**	**NC Thinking Skill(s):**
☐ Goal 1 ☐ Goal 2 ☐ Goal 3 ☐ Goal 4	_____	☐ Knowledge ☐ Generating ☐ Organizing ☐ Integrating ☐ Applying ☐ Evaluating ☐ Analyzing
Difficulty Level: ☐ Easy ☐ Medium ☐ Hard	**Item Writer**:	**Artwork required?** ☐ Yes (if checked, please attach and document source) ☐ No

Science Test Item:

Correct Answer: _____

General Guidelines:

☐ Focus directly on the objective.
☐ Write stem as a complete statement/question.
☐ Write distractors of equal length.
☐ Write distractors using same context and similar ideas.
☐ Make distractors grammatically consistent with the stem.
☐ Make each distractor plausible (and document reasoning).
☐ Avoid using negatives in distractors.
☐ Check punctuation, spelling, and grammatical structure of item.
☐ Use artwork as needed (and document source(s)).
☐ Practice fair representation in sex and race, avoiding culture-specific references.

APPENDIX E

Multiple Choice Item Writing Workbook

prepared by

Teresa Brumfield
Research Associate
Center for Educational Research & Evaluation
University of North Carolina -- Greensboro
Greensboro, NC 27402

June 2005

under CERE-UNCG subcontract with Duke University
for Teachers and Scientists Collaborating (TASC) Project

Multiple Choice Item Writing Workbook

Introduction

Thank you for your interest in serving as an item writer. The following information is intended to acquaint you with the item writing process. If you have any questions, please feel free to contact Teresa Brumfield, Research Associate, at the Center for Educational Research & Evaluation, University of North Carolina--Greensboro, 210 Curry Building, Greensboro, NC 27402, or at tebrumfi@uncg.edu.

The stated goal of the North Carolina Standard Course of Study in Science is to achieve scientific literacy, defined by the National Science Education Standards as "the knowledge and understanding of scientific concepts and processes required for scientific decision making, participation in civic and cultural affairs, and economic productivity." (p. 22). Thus, scientific literacy includes the ability to:

- find or determine answers to questions derived from everyday experiences;
- describe, explain, and predict natural phenomena;
- understand articles about science;
- engage in non-technical conversation about the validity of conclusions;
- identify scientific issues underlying national and local decisions; and
- pose explanations based on evidence derived from one's own work.

(NC SCS, 2004, p. 9)

TASC's purpose is to provide North Carolina K-8 students with opportunities to learn to think as scientists, that is, critically, creatively, and independently. To accomplish this purpose, TASC provides teacher trainings, which are focused on science content, inquiry-based teaching, and effective use of science materials. In this way, TASC trains teachers—using NSF-approved science curriculum kits—to create situations in which students take the role of scientists. That is, students observe and question phenomena, pose explanations of what they see, devise and conduct tests to support or contradict their theories, analyze data, draw conclusions from experimental data, design and build models, and discuss their findings. (Retrieved May 9, 2005, from http://tasc.pratt.duke.edu/ about.overview.php).

TASC's anticipated outcomes regarding student achievement include improving students' skills in science process and content, improving student readiness for high school science, and improving math and language arts end-of-grade test performance through inquiry-based science. TASC's anticipated outcomes regarding K-8 teachers include improvement in teachers' knowledge, attitudes, and behavior in science process and content.

To determine whether, and to what degree, these anticipated outcomes have been achieved, TASC is evaluating its program through the systematic collection and analyses of data. The Center—which has many years' experience in educational research, measurement, and evaluation—has been hired by TASC to develop student and teacher assessments designed to measure science content and process knowledge. Thus, the purpose of each test will be to assess students' and teachers' science content and process knowledge before and after the use of the particular curriculum unit for which the test is written.

NC Thinking Skills

From *Encyclopædia Britannica Online*, taxonomy is defined broadly as the science of classification or, more strictly, as biological classification. The term taxonomy has been borrowed by the field of education to describe a comprehensive, hierarchical classification scheme for instructional outcomes. A well-known taxonomy of cognitive instructional outcomes is Bloom's Taxonomy (Bloom, 1956). Another perhaps less well-known taxonomy is Marzano's Dimensions of Learning (Marzano, 1988).

The N. C. Department of Public Instruction (NC DPI) has adopted the NC Thinking Skills as their model to classify questions for North Carolina tests. The NC Thinking Skills model appears to be a blend of Bloom's Taxonomy and Marzano's Dimensions of Learning (see Table 1). Appendix A includes a table of the NC Thinking Skills, along with each category's definition, action words, and example trigger questions.

Table 1. Taxonomies used to classify instructional objectives

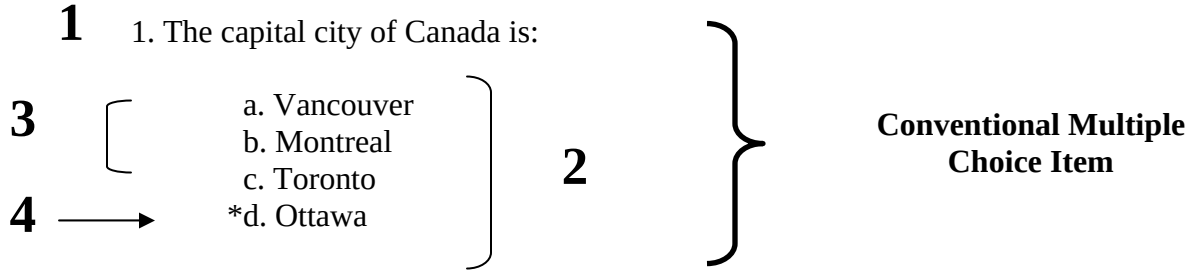
Bloom's Taxonomy	**Marzano's Dimensions of Learning**	**NC Thinking Skills**	**Description of NC Thinking Skills** (from *North Carolina Thinking Skills: An introduction* by Tom Munk (Oct. 2001)
knowledge	focusing information-gathering remembering	knowledge	"At the lowest level, students should learn the focusing, information-gathering, and remembering skills that allow them to gain declarative and procedural **knowledge**."
comprehension	organizing	organizing	"Techniques such as comparing, classifying, ordering, and representing allow students to develop skill in **organizing** information."
application		applying	"**Applying** their knowledge to a novel situation is a higher-level skill that our children will need to succeed, both in school and outside the classroom."
analysis	analyzing	analyzing	"By examining the parts and relationships of existing information, students clarify their knowledge and practice the learning skill of **analyzing**."

Bloom's Taxonomy	Marzano's Dimensions of Learning	NC Thinking Skills	Description of NC Thinking Skills (from *North Carolina Thinking Skills: An introduction* by Tom Munk (Oct. 2001)
	generating	generating	"By inferring, predicting, and elaborating, students can become skilled at **generating** new information, meaning, or ideas."
synthesis	integrating	integrating	"**Integrating** can be accomplished by condensing information efficiently into a cohesive statement or by connecting existing and prior knowledge into a new understanding."
evaluation	evaluating	evaluating	"**Evaluating** ideas by setting criteria and confirming the accuracy of claims is the last of the North Carolina Thinking Skills."

Item Classification

Since one of the outcomes of NC education is that students learn to think critically and creatively, and since the NC Standard Course of Study provides the competencies that students should demonstrate, it follows that all test questions will be classified by two dimensions:
(3) by the Instructional Objective being measured by the question, and
(4) by the NC thinking skill(s) the student will utilize to correctly answer the question.

Multiple Choice Question Terminology



1. **Stem**: presents the problem or question
2. **Options:** include the distractors (also known as "foils") and the keyed (correct) response
3. **Distractors**: appear to be reasonable answers to the examinee who does not know the content
4. **Item Key**: correct (or best) answer

The MC item stem explains the basis for answering either:

- the question to be answered, or
- the problem to be solved, or
- the incomplete statement to be completed.

The stem is followed by two or more options, which include the correct (or best) answer (i.e., Item Key) and distractors. The correct answer is the one—and only one—answer . (Jacobsen, D. M. (undated)).

Distractors are the most difficult part of the test item to write. Distractors are wrong answers but each must be plausible to test-takers who have not yet learned the knowledge that the test item is supposed to measure. To those who possess the knowledge asked for in the item, the distractors are clearly wrong choices. Distractors should resemble the correct choice in grammatical form, style, and length. After analyzing a variety of tests, Haladyna and Downing (1993) found that most items had only one or two "working" distractors leading them to conclude that three options (that is, a correct answer plus two distractors) seemed natural. (Haladyna, 1999).

Item Shells

Item writing can be a slow and arduous process. One technique that serves to accelerate the item-writing process and produce high quality items is the item shell. Item shells simplify writing items that aim to measure higher levels of cognitive behavior. (Haladyna, 1999).

An item shell is defined as a "hollow" item containing a syntactic structure that is useful for writing sets of similar items. Each item shell is a generic multiple choice test item. For example, consider the following simple item shell:

Which is an example of (any concept)*?*
A. Example
B. Plausible non-example
C. Plausible non-example

One of the limitations of the item shell technique is over-reliance upon one item shell resulting in an abundance of items with the same syntactic structure. In order to address this limitation, it is recommended that one use a variety of item shells. In this way, a variety of test questions representatively sampling the content areas for the test can be created. (Haladyna, 1999).

Another limitation of the item shell technique is that it may not be appropriate to assess all types of content and cognitive behaviors. There are instances where the learning task is specific enough so that generalization to sets of similar items is not possible. (Haladyna, 1999).

One way to develop an item shell is to use the generic shells presented in Figure 1. These are derived from item stems taken from successfully performing items. The content expert identifies the fact, concept, principle, or procedure being evaluated and the type of cognitive behavior desired (i.e., recalling, defining, predicting, evaluating, or problem-solving).

- ***Understanding – Concepts***.
  - Which is the best definition of this [concept]?
  - Which is the meaning of this [concept]?
  - Which is synonymous with this [concept]?
  - Which is like this [concept]?
  - Which is characteristic of this [concept]?
  - Which is an example of this [concept]?

- ***Understanding – Principles***.
  - Which is the best definition of ...?
  - Which statement below exemplifies the principle of ...?
  - Which is the reason for or cause of ...?
  - Which is the relationship between ... and ...?
  - Which is an example of the principle of ...?

- ***Critical Thinking – Predicting Using a Principle***.
  - What would happen if ...?
  - If (there is an action), then what happens?
  - What is the consequence of [an action]?
  - What is the cause of a [result]?
  - [Information given.] What is the expected result?
  - Which distinguishes [one concept from another concept]?

- ***Critical Thinking – Evaluating Using Facts and Concepts***.
  - Which is the [most or least] [important, significant, effective] ...?
  - Which is [better, worse, higher, lower, farther, nearer, heavier, lighter] ...?
  - Which is [most like, least like] ...?
  - What is the difference between ... and ...?
  - What is a similarity between ... and ...?

- ***Critical Thinking – Evaluating Using a Principle***.
  - Which of the following principles applies to evaluating ...?
  - What is the most important factor contributing to ...?

- ***Evaluating – Procedures***.
  - Which of the following procedures best applies to the solution of [a problem]?

- ***Problem Solving – Concepts, Principles, Procedures***.
  - [Problem presented.] What is the best way to solve [this problem]?
  - [Problem presented.] What is the solution?

Figure 1. Generic item shells (Haladyna, 1999).-

Another way to develop an item shell is to transform highly successful items into item shells. As an illustration, consider an item shell for an eighth grade science unit on gases and their characteristics:

- Step 1: Identify the stem.

  Which is the distinguishing characteristic of hydrogen?

- Step 2: Underline the key word or phrase.

  Which is the distinguishing characteristic of <u>hydrogen</u>?

- Step 3: Identify variations for each key word or phrase.

  Which is the distinguishing characteristic of
  [any gases studied in this unit]?

- Step 4: Select an instance from the range of variations.

  Oxygen.

- Step 5: Write the stem.

  Which is the distinguishing characteristic of oxygen?

- Step 6: Write the correct answer.

  A. It is the secondary element in water.

- Step 7: Write the distractors.

  B. It has a lower density than hydrogen.
  C. It can be fractionally distilled.
  D. It has a lower boiling point than hydrogen.

As illustrated, the last word in the stem can be replaced by any of a variety of gases, easily producing many item stems. However, the difficult task of choosing a right answer and several plausible distractors remains. Even so, the value of the item shell is its versatility to operate at different cognitive levels with the four types of content (facts, concepts, principles, and procedures) and in different subject matter areas. (Haladyna, 1999).

Multiple Choice Item Writing Guidelines

General

1) **<u>Base each item on specific content and type of mental behavior</u>**. (Haladyna, 1999.) That is, an item should assess *one* instructional objective although more than one NC thinking skills could be needed to answer the question correctly.

   Figure 2 is an example of a fourth grade science item developed by NC DPI.

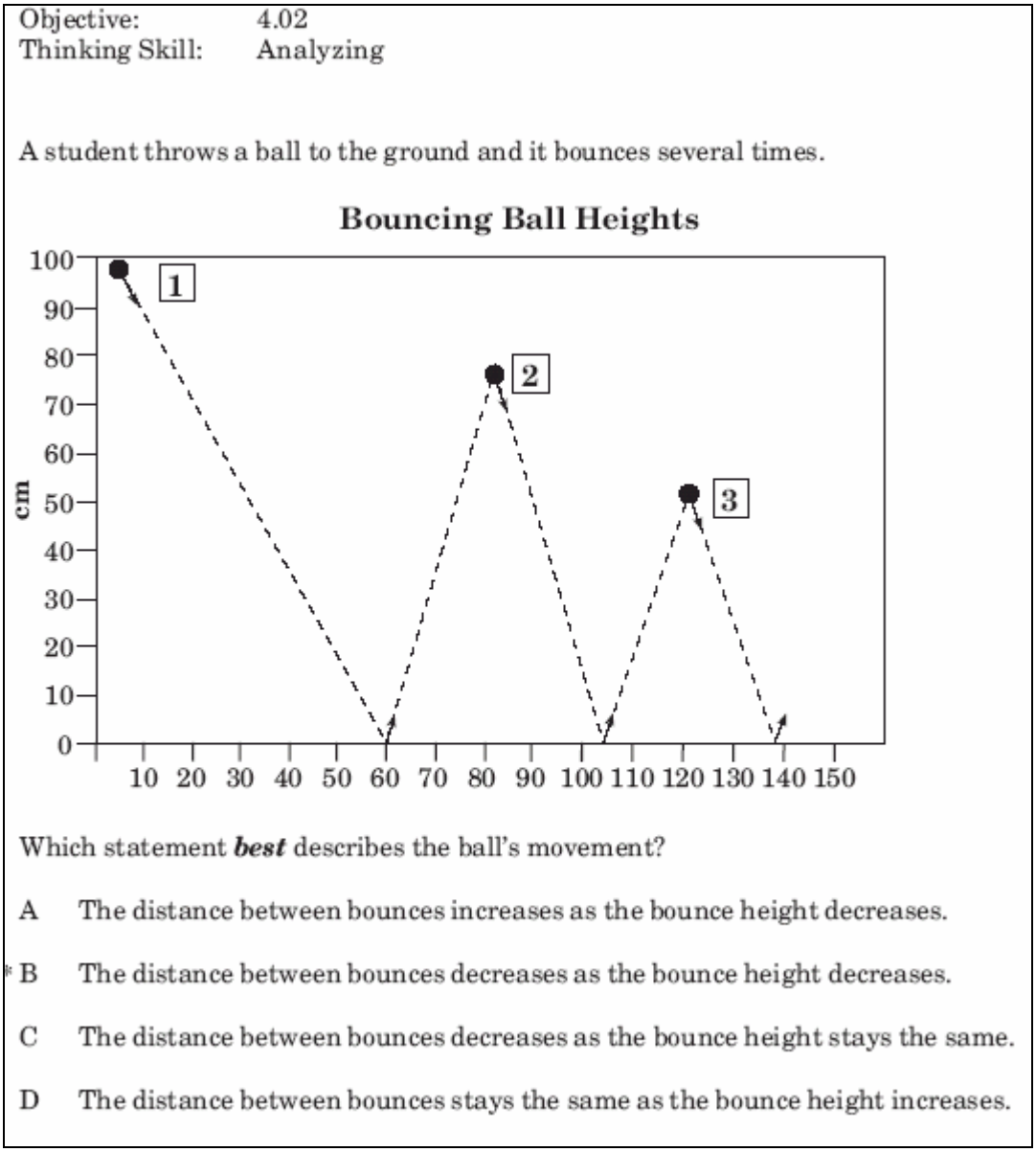


Figure 2. Grade 4 science item (NC DPI)

2) **<u>Each item should be as short and verbally uncomplicated as possible</u>**.
   a) Write the item at a reading level that is appropriate for all students being tested. (Haladyna, 1999).
   b) Give as much content as is necessary to answer the question but avoid superfluous information. (Cohen & Wollack, 2003; Haladyna, 1999).
   c) Use clear and concise directions for each item. Directions should specify the task for examinees by defining the activity required and focusing attention on relevant materials. (Osterlind, 1998).

   For example:

   Read the passage below and answer the question that follows.

   Use the graph below to answer the question that follows.

   Use the diagrams and the paragraph below to answer the question that follows.

3) **<u>Keep the specific content of items independent from one another</u>**. That is, avoid providing information in one item that cues the testwise student to the correct answer in another item. (Haladyna, 1999.)

   For instance, consider a line of questions focusing on main ideas of a novel. After answering item 1 correctly, this testwise student will look for clues in the next item. If "Roxie" is correct for Item 1, it must be incorrect for Item 2. Testwise students use these types of strategies to select answers to items they do not know. (Haladyna, 1999).

   The following questions come from the story "Stones from Ybarra".

   1. Who was Lupe's best friend?
      A. Kate
      B. Dolores
      C. Roxie*

   2. Who was quarreling with Lupe?
      A. Kate
      B. Sarah*
      C. Roxie

4) **Avoid opinion-based items**. Items should reflect well-known and publicly supported facts, concepts, principles, and procedures. (Haladyna, 1999.)

The item below would be a good basis for discussion but probably should not be included in an examination.

The most serious aspect of the energy crisis is the

1. *possible lack of fuel for industry.*
2. *possibility of widespread unemployment.*
3. *threat to our environment from pollution.*
4. *possible increase in inflation.*
5. *cost of developing alternate sources of energy.*

Such an item might be rewritten to focus on a more specific aspect of the energy crisis. It might also be written to focus on the opinion of a recognized expert:

According to Professor Koenig, the most serious aspect of the energy crisis is the

1. *possible lack of fuel for industry.*
2. *possibility of widespread unemployment.*
3. *threat to our environment from pollution.*
4. *possible increase in inflation.*
5. *cost of developing alternative sources of energy.*

(Retrieved April 27, 2005 from http://www.msu.edu/dept/soweb/writitem.html).

5) **Avoid trick items**. This would include items where the item writer's intention appeared to deceive, confuse, or mislead the test-taker. (Haladyna, 1999.)

For example:
Some months have 31 days. How many have 28?
(All months have 28 days.)

6) **Use correct grammar, correct punctuation, capitalization, and spelling**. That is, edit and proof items. (Haladyna, 1999.)

7) **Practice fair representation in sex and race carefully avoiding culture-specific references**. (NC Department of Public Instruction).

The Stem

8) **<u>Ensure that the directions in the stem are very clear</u>**. (Haladyna, 1999.)

The item stem should always phrase the problem to be answered by each option in a clear, unambiguous manner such that the test-taker should know what is being asked in the item.

For example, the following stem:

Regarding gravitation:

is made more complete and precise by the following revision:

Which of the following best shows the concept of gravitation?

(Haladyna, 1999).

9) **<u>Word the stem positively; avoid negatives such as NOT or EXCEPT</u>**. (Haladyna, 1999.)

A major problem with a negatively stated item is that students may miss the negation when reading the stem. A negatively stated item does require an examinee to switch his or her mind set from that of looking for the best answer to that of locating the most definite non-answer. Items with negatively stated stems can often be rewritten as effective positively-stated items.

For example, the negatively stated item

Which of the following is NOT a method of determining test reliability?

1. *Coefficient of equivalence*
2. *Coefficient of stability*
3. *K-R #20*
4. *Split-halves procedure*
5. *Test-criterion intercorrelation*

may be rephrased as a positively stated item.

Which of the following is a method of determining the validity of a test?

1. *Coefficient of equivalence*

2. *Coefficient of stability*
3. *K-R #20*
4. *Split-halves procedure*
5. *Test-criterion correlation*

The correct answer to each of the two above items is option 5.

(Retrieved April 27, 2005 from http://www.msu.edu/dept/sweb/writitem.html).

10) **Avoid excessive verbiage**. (Haladyna, 1999.)

For instance, the following item is unnecessarily verbose:

When a police officer arrests someone, the officer must inform the person of certain rights including the right to remain silent and the right to an attorney. Why is this required?

A. *Criminal law places the burden of proof on the defendant.*
B. *The rights of suspects vary in different areas of the country.*
C. *Many people who are arrested tend to be poorly educated.*
D. *The government is required to respect citizens' rights.*

A more concise wording of this item would be:

Why must police officers inform suspects of certain rights during an arrest?

A. *Criminal law places the burden of proof on the defendant.*
B. *The rights of suspects vary in different areas of the country.*
C. *Many people who are arrested tend to be poorly educated.*
D. *The government is required to respect citizens' rights.*

(Retrieved April 27, 2005 from http://www.mdk12.org/instruction/curriculum/hsa/government/common_mistakes.html).

11) **<u>Include the central idea in the stem, instead of the choices</u>**. (Haladyna, 1999.) That is, state the problem as a complete statement or question.

The following item

Multiple-choice items

1. *may have several correct answers.*
2. *consists of a stem and some options.*
3. *always measure factual details.*

does not have a problem or question posed in the stem. The examinee cannot determine the problem on which the item is focused without reading each of the options. The item should be revised, perhaps to read

The components of a multiple-choice item are a

1. *stem and several foils.*
2. *correct answer and several foils.*
3. *stem, a correct answer, and some foils.*
4. *stem and a correct answer.*

A student who has been given the objective of recognizing the components of a multiple-choice item will read the stem and immediately know the correct answer. The only remaining task is to locate the option which contains the complete list of components. (Retrieved April 27, 2005 from http://www.msu.edu/dept/soweb/writitem.html).

The Alternatives

12) **<u>Use as many good choices as possible</u>**, but three seems to be a natural limit. (Haladyna, 1999.)

13) **<u>Make sure that only one of these choices is the correct answer</u>**. (Haladyna, 1999.)

14) **Place choices in logical or numerical order**. (Haladyna, 1999.)

If a student who understands the principle being examined determines the correct answer after reading the item stem, then s/he should not have to spend time searching for that answer in a group of haphazardly arranged options. Options should always be arranged in some systematic manner, e.g., dates arranged chronologically, numerical quantities in ascending order of size, and names in alphabetic order. Consider the following example.

What type of validity is determined by correlating scores on a test with scores on a criterion measured at a later date?

1. *Concurrent*
2. *Construct*
3. *Content*
4. *Predictive*

A student properly recognizing the description of predictive validity in the stem of the above item may go directly to the correct option since the options are in a logical order.

(Retrieved April 27, 2005 from http://www.msu.edu/dept/soweb/writitem.html).

15) **Keep choices independent; choices should not be overlapping**. (Haladyna, 1999.)

A knowledgeable examinee must be able to locate only one option which will contain the correct or best answer.

Consider the faulty item below.

What should be the index of difficulty for an effective mastery-model test item?

1. *Less than 10*
2. *Less than 20*
3. *More than 80*
4. *More than 90*

If the index of difficulty is expressed as the proportion of the examinees who answer an item correctly, and option 1 is correct, then option 2 is also correct. The item should be rewritten as follows:

What should be the index of difficulty for an effective mastery-model test item?

1. *Approximately 10*
2. *Approximately 20*
3. *Approximately 80*
4. *Approximately 90*

(Retrieved April 27, 2005 from http://www.msu.edu/dept/soweb/writitem.html).

16) **Keep options homogeneous in content**. (Haladyna, 1999.)

The use of options that are heterogeneous in content often cues students to the correct answer. Thus, a standard practice of keeping options homogeneous avoids the possibility of giving away the correct answer. The following item illustrates both homogeneous and heterogeneous options:

HOMOGENEOUS OPTIONS	HETEROGENEOUS OPTIONS
What will make salsa hottest? A. Adding habanero chili peppers B. Adding Anaheim chili peppers C.* Adding jalapeno chili peppers	What will make salsa hottest? A. Adding the seeds of peppers B. Adding spices C.* Adding jalapeno chili peppers

(Haladyna, 1999).

17) **Keep the length of options about equal**. (Haladyna, 1999.)

If a test writer consistently writes correct options which are of different length than the foils or distracters, students will quickly learn to select correct answers on the basis of these idiosyncrasies. Longer correct options are perhaps most common since it is often necessary to add qualifiers to allow an option to be correct.

For example:

A random sample is one in which

1. *subjects are selected by levels.*
2. *each subject has an equal probability of being chosen for the sample.*
3. *every nth subject is chosen.*
4. *groups are the unit of analysis.*

The item might be rewritten:

A random sample is one in which

1. *subjects are selected by levels in proportion to the number at each level in the population.*
2. *each subject has an equal probability of being chosen.*
3. *every nth subject is chosen from a list.*
4. *groups, rather than individuals, are the unit of analysis.*

In the above revision, the correct option 2 is not conspicuously longer, as it was in the original version. In any case, shorter or longer correct options are not a problem unless they are consistently shorter or longer, so that students may establish a rule.

(Retrieved April 27, 2005 from http://www.msu.edu/dept/soweb/writitem.html).

18) **Avoid using *none-of-the-above*, *all-of-the-above*, or *I-don't-know***. (Haladyna, 1999.)

19) **Avoid giving clues to the right answer, such as**:

a) **Specific determiners including *always*, *never*, *completely*, and *absolutely***.

Specific determiners—i.e., *always, never, totally, absolutely,* and *completely*—are so extreme that they are seldom the correct answers. When a specific determiner *is* the right answer, its use is justified *if* the distractors also contain other specific determiners.

Which of the following is most likely to produce the most student learning?
A. Never assign homework on Fridays.
B. Homework is consistent with class learning.*
C. Always evaluate homework the next day.
(Haladyna, 1999).

b) **Options with words or phrases identical to, or resembling, words in the stem**. (Haladyna, 1999).

For example:

Who were known as the Magnificent Seven?
A. A touring softball team

B. A group of seven cowboys*
C. A rock and roll group

c) **<u>Grammatical inconsistencies that cue the test-taker to the correct choice</u>**. (Haladyna, 1999).

Students are quick to take advantage of extraneous clues such as inconsistent stem and options. Thus they are responding to the item in terms of verbal skills possibly quite different from the skills the item is intended to measure. Note the extraneous clues in the item below.

A test which can be scored by a clerk untrained in the content area of the test is an

1. *diagnostic test.*
2. *criterion-referenced tests.*
3. *objective test.*
4. *reliable test.*
5. *subjective test.*

The examinee is led directly to option 3 by the last word in the stem which requires an option with its first word beginning with a vowel. Option 2 is rendered more implausible by the singular- plural inconsistency. The item might be rewritten as follows:

A test, which can be scored by a clerk untrained in the content area of the test, is said to be

1. *diagnostic.*
2. *criterion-referenced.*
3. *objective.*
4. *reliable.*
5. *subjective.*

(Retrieved April 27,2 005 from http://www.msu.edu/dept/soweb/writitem.html).

d) **<u>Blatantly absurd, ridiculous options</u>**. (Haladyna, 1999.)

While it may be tempting to create a ridiculous option and/or a humorous option—especially when writing that third or fourth option, these options will seldom be selected. (Haladyna, 1999).

20) **Make all distractors plausible**. (Haladyna, 1999.)

The major purpose of a multiple-choice item is to identify examinees who do not have complete command of the concept or principle involved. In order to accomplish this purpose, the distracters must appear as reasonable as the correct answer to students who have not mastered the material. Consider the following item:

A terminal may be defined as

1. *a final stage in a computer program.*
2. *the place where a computer is kept.*
3. *an input-output device used when much interaction is required.*
4. *an auxiliary memory unit.*
5. *a slow but simple operating system.*

Options 1 and 2 are derived from the common use of the word "terminal." They were each chosen by a number of students when the item was used in a pretest. Option 3 was keyed as the correct option.
(Retrieved April 27, 2005 from http://www.msu.edu/dept/soweb/writitem.html).

In writing multiple choice questions, good, plausible distractors come from a thorough understanding of students' common or typical errors. (Haladyna, 1999.)

Criteria for Evaluating Multiple Choice Items

Table 2. Multiple choice item writing checklist

	Yes	No
4. General		
a. Is the wording of the item clear and unambiguous?		
b. Does the item present one--and only one--problem?		
c. Is the item written at appropriate reading level for all students?		
d. Does each item measure only one instructional objective?		
e. Have punctuation, capitalization, spelling, and grammatical structure of the item been checked?		
f. Does the item avoid culture-specific references?		
5. The Stem		
a. Is the problem stated concisely as a complete statement/ question?		
b. Is the stem presented positively?		
c. Are the directions in the stem clearly stated?		
d. Have extraneous cues to the correct answer been avoided?		
6. The Alternatives		
a. Is there one--and only one--clearly correct answer?		
b. Is the correct answer supported by documentation (and not an expression of opinion)?		
c. Are the incorrect alternatives logical and plausible and unlikely to be eliminated by someone who does not know the material?		
d. Are the alternatives grammatically consistent with the stem?		
e. Is the correct response about the same length as one or more of the distractors and not any more technical than the other responses?		
f. Have *none-of-the-above, all-of-the-above,* or *I-don't-know* been avoided as alternatives?		

References

Cohen, A. S. and Wollack, J. A. (2003). *Handbook on test development: Helpful tips for creating reliable and valid classroom tests*. Retrieved April 27, 2005 from http://wiscinfo.doit.wisc.edu/exams/Handbook%20on%20Test%20Construction.pdf.

Encyclopædia Britannica. Retrieved May 20, 2005, from Encyclopædia Britannica Online. <http://search.eb.com/eb/article?tocId=9110579>

Haladyna, T. M. (1999). *Developing and validating multiple-choice test items: 2nd edition.* Mahwah, NJ: Lawrence Erlbaum Associates.

Jacobsen, D. Michele (undated). "Multiple Choice Item Construction" Powerpoint presentation. Retrieved April 19, 2005 from http://www.ucalgary.ca/~dmjacobs/portage/.

Marzano, R. J., Brandt, R. S., Hughes, C. S., Jones, B. F., Presseisen, B. Z., Rankin, S.C. & Suhor, C. (1988). *Dimensions oftThinking: A framework for curriculum and instruction.* Alexandria, VA: Association for Supervision and Curriculum Development.

Michigan State University (undated). *Writing test items*. Retrieved April 27, 2005 from http://www.msu.edu/dept/soweb/writitem.html.

Munk, T. (2001). *North Carolina Thinking Skills: An introduction*. Retrieved April 19, 2005 from http.//www.learnnc.org/.

NC Department of Public Instruction (2004). *Science: Standard course of study and K-12 grade level competencies*. Raleigh, NC: NC DPI.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats, 2nd edition*. Boston: Kluwer Academic Publishers.

School Improvement in Maryland (undated). *Common mistakes in item writing.* Retrieved April 27, 2005 from http://www.mdk12.org/instruction/curriculum/hsa/government/common_mistakes.html.

Taxonomy of educational objectives, Handbook I: Cognitive domain by B. S. Bloom (Ed.), 1956, New York: David McKay.

Appendix A

Category	Definition	Action Words	Example	Examples of Trigger Questions
Knowledge	Most tasks require that learners recognize or remember key facts, definitions, concepts, rules, and principles. When content is new, students must be guided in relating the new knowledge to what they already know, organizing and then using that new knowledge. Knowledge questions require students to repeat verbatim or to paraphrase given information. To know information, students need most often to rehearse or practice it, and then to associate it with other, related concepts. The Bloom taxonomy levels of knowledge and comprehension are subsumed here, since verbatim repetition and translation into the student's own words represent acceptable evidence of learning and understanding.	• Define • Repeat • Identify • What • Label • When • List • Who • Name	List the names of the main characters in the story.	• Define the word *xxxxx* . • What is a *xxxxxxxx*? • Label the following. • Identify the *xxxxxxx* in this *yyyyyy* . • Who did *xxxxxxx* ?

Category	Definition	Action Words	Example	Examples of Trigger Questions
	Knowledge can be of two types: Declarative (i.e., attributes, rules) or procedural (skills and processes). Items of this type are factual and content-specific.			
Organizing	This category relates to some of the skills in the Bloom level of comprehension and analysis. These tasks require learners to structure information so that it can be more deeply understood or presented more clearly. For instance, such tasks may include: • Comparing entities, identifying similiarities and differences between them. • Classifying groups of items into categories on the basis of attributes. • Ordering sequences or ordering entities acccording to a given criterion. • Representing changes in the form of the information to show how	• Compare • Differentiate • Contrast • Order • Classify • Distinguish • Relate	• Compare the properties of objects or events. • Compare the themes of these two stories. • Classify the causes and effects of separate events into categories. • Represent visually, verbally and with symbols the social political and economic, characteristics of Western Europe. • Order the settings of six novels based upon the level of details of the described scenes.	• Compare the *xxxxx* before and after *yyyyy*. • Contrast the *xxxxxx* to the *yyyyy*. • Differentiate between *xxxxx* and *yyyyy*. • Classify *xxxxx* by *zzzzzz*. • Order *zzzzz* by *xxxxx*. • Group these *xxxx* by *yyyy*.

Category	Definition	Action Words	Example	Examples of Trigger Questions
	critical events are related (visual, verbal, and symbolic). Higher levels of organizing include grouping items into categories based on their features, sequencing things according to a given characteristic, and representing by changing the form of the information to show relationships, such as taking and understanding a text and explaining things visually.			
Applying	Application is based on a learner's ability to apply prior learning to a new or a novel situation without having to be shown how to do so. The task is to bring together the appropriate information, generalization, or principle (declarative and procedural knowledge) that are required to solve a problem. Thus teachers should create novel situations and expect	• Apply • Demonstrate • Calculate • Complete • Illustrate • Show • Solve • Examine • Modify • Relate • Change • Classify • Experiment • Discover • Dramatize • Sketch	• Apply your knowledge of swimming and weight lifting to create a new sports game for 5th graders. • Apply your knowledge of spreadsheets, mathematics, and the planets we have been studying to create a spreadsheet that calculates how much	(It is not the style of question that is important, but that the question apply previously taught and learned information to a novel situation.)

Category	Definition	Action Words	Example	Examples of Trigger Questions
	learners to apply prior knowledge to higher order tasks without being shown what to do. That is, the learners must recognize when information or skill are needed and then use them to solve new problems or completely novel tasks.		each person will weigh on each of the planets in our solar system. • Demonstrate using these objects the orbit of a planet that orbits around two stars instead of one.	
Analyzing	In this operation, students divide a whole into component elements. Generally the part/whole relations and the cause/effect relationships that characterize knowledge within subject domains are essential components of more complex tasks. The components can be the distinctive characteristics of objects or ideas, or the basic actions of procedures or events. This definition of analysis is the same as that in the Bloom taxonomy. Analyzing clarifies existing information by discovering and	• Subdivide • Categorize • Break down • Sort • Separate	• **Science analysis** --Separate the components of the process. --Identify the features of animate and inanimate objects. • **Social science analysis** Analyze components or elements of an event. • **Literature analysis** Identify components of literary, expository, and persuasive discourse.	• What are the basic elements (ingredients) in a *xxxxxxx*. • What is/are the functions of *xxxxxxx*. • Inventory the parts of *xxxxxxx*. • Categorize the *xxxxxxx* of *yyyyyyyy*. • Sort the *xxxxxx*. • What is the order of steps in *xxxxxxx*.

Category	Definition	Action Words	Example	Examples of Trigger Questions
	examining parts/ relationships: • Identifying attributes and components refers to recognizing and articulating the parts that together constitute a whole. • Identifying relationships and patterns refers to recognizing and articulating the interrelationships among components (causal, hierarchical, temporal, spatial, correctional, or metaphorical; equivalence, symmetry, and similarity; difference, contradiction, and exclusion).			
Generating	Generating builds a structure of ideas that pulls together new and old information. Both deductive and inductive reasoning fall in this category. In deductive tasks, students are given a generalization and are required to	• Deduce • Anticipate • Predict what if • Infer • Apply • Speculate • Conclude	• **Science/social science** o Draw conclusions; make predictions; pose hypotheses, tests, and explanations.	• Hypothesize what will happen if *xxxxx* . • Predict what would be true if *xxxxx* . • Conclude what the result will

Category	Definition	Action Words	Example	Examples of Trigger Questions
	recognize or explain the evidence that relates to it. In inductive tasks, students are given the evidence or details and are required to come up with the generalization. Generating constructs a framework of ideas that holds new and old information together. The step of inference could also be seen as the first step of what Bloom called synthesis or Marzano called integrating. • *Inferring* refers to going beyond the available information to identify what reasonably may be true. • *Predicting* refers to assessing the likelihood of an outcome based on prior knowledge of how things usually turn out. • *Elaborating* involves adding details, explanations, examples, or other relevant		o Predict, hypothesize, and conclude. • **Literature** Infer characters' motivation; infer cause and effect.	be if *xxxxx* . • What if *xxxxx* had happened instead *yyyyy*?

Category	Definition	Action Words	Example	Examples of Trigger Questions
	information from prior knowledge in order to improve understanding (explanations, analogies, and metaphors).			
Integrating	Integrating connects or combines prior knowledge and new information to build new understandings. Thus the learner uses old ideas to create new ones, generalizes from given facts, and relates knowledge from several areas thereby demonstrating the ability to combine elements into a pattern not clearly there before. Bloom called this synthesis. • Summarizing refers to combining information effectively into a cohesive statement. It involves condensing information, selecting what is important (and discarding what is not), and	• Combine • Integrate • Modify • Rearrange • Substitute • Plan • Create • Design • Invent • What if? • Compose • Formulate • Prepare • Generalize • Rewrite • How would you test • Propose an alternative • Compose • Design • State a rule • Theorize • Develop • Devise • Originate • Revise • Extend • Synthesize • Conceive • Project • Hypothesize	Design an airplane model that flies as well upside down as right side up.	• Using *xxxxx*, how many ways can you think of to *yyyyy*? • Summarize in your own words the story of *xxxxx*. • Make a plan to *zzzzz*? • What might happen if *xxxxx*? • Can you make a *yyyyy*? • How can you improve or make *xxxxx* better? • What ideas do you have for changing *xxxxx*?

Category	Definition	Action Words	Example	Examples of Trigger Questions
	combining logical text proportions. • Restructuring refers to changing existing knowledge structure to incorporate new information. New information and prior knowledge are connected, combined and incorporated into a new understanding.			
Evaluating	These tasks require us to judge quality, credibility, worth, and/or practicality of ideas. Generally, we expect students to use established criteria and explain how these criteria are or are not met. Criteria are standards, rules, or tests on which a judgment or decision can be based. The criteria might be established rules of evidence, logic, or shared values. Bloom's levels of synthesis and evaluation are involved in this	• Evaluate • Argue • Judge • Recommend • Assess • Debate • Appraise • Critique • Defend	• Evaluate soundness and significance of findings • Evaluate credibility of arguments, decisions, and reports; evaluate significance • Evaluate form, believability, significance, completeness, and clarity	• What you would do if xxxxx happened. Why? • Judge what would be the best way to solve the problem of xxxx .. • Why did you select that solution? • Evaluate whether you would xxxxx or yyyyy . • Xxxxx in this

Category	Definition	Action Words	Example	Examples of Trigger Questions
	category. To evaluate, students must assemble and explain the interrelationship of evidence and reasons in support of their conclusion (synthesis). Explanation of criteria for reaching a conclusion is unique to evaluative reasoning. • *Establishing criteria* sets standards for judging the value or logic of ideas. • *Verifying* refers to confirming or proving the truth of an idea, using specific standards or criteria of evaluation (checking the accuracy of facts, checking the meaning or accuracy of the author's statement by looking back at the text, using research results to verify the hypotheses).			situation. Why?